Beyond the conventional information limit: the relevant coherence function

P.D. Nellist, J.M. Rodenburg

Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge CB3 0HE, UK

(Received 30 July 1993; in final form 28 December 1993)

Abstract

The theory of partial coherence functions as applied to a super-resolution reconstruction algorithm is developed in detail, taking into account the phase gradients across the aperture function, the source and detector sizes, and temporal coherence. Experimental examples demonstrate the main properties of the relevant coherence envelopes, and why they need not limit the total band-pass of the microscope. It is demonstrated that finite detector size in the STEM configuration can facilitate a simpler reconstruction algorithm by creating a virtual objective aperture which obviates the need for a physical objective aperture. These results are also relevant to the question of uniquely deconvolving shadow images.

1. Introduction

We show theoretically and experimentally how it is possible to separate all the incoherent contributions which arise in the scanning transmission electron microscope (STEM) in order to increase the bandwidth of a high-energy electron scattering experiment well beyond the confines of the conventional "information limit", in a way which is robust and fully quantifiable. Konnert et al. [1] have already shown that very-high-resolution information can be extracted from the dark-field scattered intensity in the coherent microdiffraction pattern of a STEM, when the data is also collected as a function of the probe position, and processed by first taking a Fourier transform over the detector plane. The resulting data set consists of a position-resolved Patterson function, in which it is possible to seek the mid-points of pairs of scattering centres or atom columns. In this paper, we employ a more extensive data set, including also the unscattered zero-order disc (or Gabor hologram), and consequently choose to process the data in a different way. It should be pointed that we do not seek to process either diffraction patterns or images, but a four-dimensional data set containing all this information simultaneously. Our method follows earlier work [2–5] in which it was shown that by taking a Fourier transform with respect to both the diffraction coordinate and the probe-position coordinate, the complex transfer of the lens and partial coherence function of the STEM source become completely separable. Further experience with this technique [6] has shown that it is the Fourier transform with respect to the probe position which most usefully recasts the measured data into its most separable form.
Having first taken this transform, we may then observe various sections through the resultant four-dimensional data set, and it is found that all the various incoherence terms become obvious. In theory, this form of aperture synthesis requires negligible source coherence. Indeed, it has been shown on the optical bench \[7\] that for thin crystalline specimens, in which case the reconstruction algorithm reduces to a form of coherent crystallography which simply has to determine the phase of a handful of reflections, it is possible to generate very good images at very many times the conventional resolution limit, even if the incoherent source size is a significant proportion of the unit cell size. However, in order to achieve these sorts of gains in resolution, it is necessary to perform a process which we call “stepping out” into reciprocal space. That is to say, the first-order diffraction discs are phased with respect to the zero-order diffraction disc in the convergent beam electron diffraction (CBED) pattern, and then second-order diffraction discs are phased with respect to the first order and so on and so forth, up to whatever resolution is required. There are to disadvantages with this method. (i) Although it can deal with strong objects, it breaks down, at least in its simple form, for objects of substantial thickness (i.e. those in which significant Fresnel propagation occurs within the specimen) because each “step out” corresponds to a scattering experiment performed at a slightly different angle, and so the reconstruction thus obtained is not easy to interpret. (ii) In the more general case of an aperiodic or finite object, it is necessary to collect and store a very large data set and deconvolve using a Wigner distribution. In practical terms, this means we must also Fourier transform the data with respect to the diffraction coordinate (see, for example, the optical experiments \[4,3\]) and so this process is computationally very intensive.

A much simpler scheme involves obtaining only double resolution information by processing only the zero-order disc as a function of the Fourier transform of the probe position \[6\]. The absence of “stepping out” has two important advantages. (i) Consideration of the three-dimensional scattering geometry shows that such a method picks out exactly a two-dimensional plane in reciprocal space, corresponding to a perfect projection of the specimen in real space: a property which has been recently demonstrated on the optical bench \[8,9\]. Because excitation errors due to the curvature of the Ewald sphere are suppressed, this would imply that a specimen tilt series tomographic reconstruction at sub-Ångström resolution may become feasible. (ii) The computing involved is relatively small – it requires each detector element in the coherent CBED plane of a STEM to pick out only one Fourier component of the probe position coordinate – a calculation which could be straightforwardly hard-wired for fast, large field of view imaging.

The double resolution reconstruction may at first appear to depend upon the existence of an objective aperture in the back focal plane, which allows clean separation of the two side-bands (as discussed in Ref. \[6\]). However, in the same way as beam divergence limits the coherence envelope in CTEM, the detector pixel size in STEM creates a similar virtual objective aperture, as demonstrated below. A mundane but important advantage of this effect is that if we do not require an objective aperture, the specimen contamination problems inherent in STEM are much less severe, because a large area is being irradiated, in a way similar to the conventional transmission microscope. It means also that the same processing method could be applied to shadow image microscopy \[10\].

These results can form a useful basis for specifying the optimal characteristics for a new generation of electron microscope which will not rely on extending the first crossing point of the contrast transfer function, or even on the total width of the coherence envelope. Instead, we may re-formulate the imaging problem as an inverse aperture-synthesis calculation which should be made as cheap and as fast as possible. The important constraint is therefore the number of numbers we must measure, either in diffraction space or image space. We show below that these figures are indirectly affected by the form of the lens transfer function, or, in the case of shadow image microscopy, the form of the propagator lying between the source and the specimen.
2. The phase-retrieval / super-resolution algorithm

In what follows we denote reciprocal-space vectors by a prime, so \( q' \) is conjugate to the real-space vector \( q \). Reciprocal space functions are capitalised and are related to their real-space version by the Fourier transform,

\[
F(q') = \int f(q) \exp(i2\pi q \cdot q') \, dq.
\]  

The phase-retrieval / super-resolution algorithm operates upon a data set, which we call \( |M(r', p)|^2 \), which is the intensity of the CTEM image recorded as a function of illumination angle, or equivalently by reciprocity \([11]\) the intensity of the STEM microdiffraction pattern recorded as a function of probe position. The reciprocal-space vector \( r' \) indentsifies the CTEM illumination angle, or the position in the STEM microdiffraction plane, and \( p \) is a real-space vector denoting the position in the CTEM image plane, or the STEM probe position. Vectors \( q \) and \( q' \) are used as general vectors when a reference to neither probe position nor position in the microdiffraction plane is implied. If we assume the specimen is thin (but not necessarily weak) so that it can be described by a complex transmission function \( t_{\text{in}}(q) \), which is what the exit wave function of the specimen would be under conditions of parallel illumination, then it can be shown that if we take a Fourier transform of \( |M(r', p)|^2 \) with respect to the probe position, then we obtain a quantity we call \( G(r', p') \), such that

\[
G(r', p') = G_A(r', p') \ast D(r', p'),
\]

where \( \ast \) denotes a convolution with respect to the \( r' \) coordinate only. The function \( D(r', p') \) is given by

\[
D(r', p') = \Psi(r') \Psi^*(r' - p'),
\]

where \( \Psi(q') \) is the Fourier transform of \( \psi(q) \), and \( G_A(r', p') \) is given by

\[
G_A(r', p') = A(q') A^*(r' + p'),
\]

where \( A(q') \) describes both the modulus and phase of the objective lens transfer function. Of course, the complete separation of the aperture function from \( D(r', p') \) requires us to take a further Fourier transform with respect to \( r' \), to form a quantity we call \( H(r', p') \), as described in detail elsewhere \([3]\). However, here we limit ourselves only to this first Fourier transform with respect to \( p \), because it is most easy to discuss the relevant incoherence envelopes by considering how they attenuate the function \( G_A(r', p') \). Furthermore, for many classes of specimen, particularly crystalline specimens and weak objects, most pertinent structural information is automatically separated in \( G(r', p') \), as we show experimentally below. The aperture function can be further decomposed into its modulus component (a physical aperture), and its phase component, \( \chi \), where in the absence of astigmatism \( \chi(q') \) may be written

\[
\chi(q') = \pi z \lambda |q'|^2 + \frac{1}{2} \pi C_s \lambda^3 |q'|^4
\]

where \( z \) is the defocus of the lens, \( C_s \) is the coefficient of spherical aberration and \( \lambda \) is the electron wavelength. \( G(r', p') \) can be thought of as a set of diffractograms of CTEM images with illumination tilt \( r' \).

\( G_A(r', p') \) is a four-dimensional function, so it is only possible to give diagrams of planes extracted from it. Fig. 1a shows the magnitude of a plane in \( G_A(r', p') \) plotted as a function of one component of \( r' \) and the parallel component of \( p' \) for an objective aperture of radius \( C_C \). It is important to note that for \( p' = -2r' \) the phase of \( G_A(r', p') \) is zero if \( \chi(q') \) is an even function \( q' \), such as is the case for spherical aberration, defocus and astigmatism. Fig. 1b shows the magnitude

![Fig. 1. Regions of unity magnitude in \( G_A(r', p') \): (a) along corresponding parallel components of \( r' \) and \( p' \), the other components of both \( r' \) and \( p' \) are zero; (b) as a function of \( r' \) for constant \( p' \); (c) as a function of \( p' \) for constant \( r' \).](https://example.com/fig1.png)
Fig. 2. Region of significant magnitude in \( G(r', \rho') \) along corresponding parallel components of \( r' \) and \( \rho' \); the other components of both \( r' \) and \( \rho' \) are zero.

The easiest way to demonstrate the nature of Eq. (2) is by some experimental data, collected from a VG HB501 STEM (100 keV, \( C_s = 3.1 \) mm). Fig. 3 is an incoherent CBED pattern taken from graphite (in practice we have obtained this by scanning a coherent probe over the specimen, and summing everything which arrives in the coherent microdiffraction plane). Fig. 4 is a single coherent microdiffraction pattern – note the usual interference fringes in the region of disc overlap (as described in Ref. [12]). Now we scan the probe over the specimen and form \( G(r', \rho') \) by taking the Fourier transform with respect to the probe position coordinate. One finds that the data set has little amplitude except at values of \( \rho' \) corresponding exactly to the frequency (as a function of probe position) at which the interference fringes in Fig. 4 switch on and off, as in STEM lattice imaging [13]. In Fig. 5 we show this slice (at constant \( \rho' \), plotted as a function of \( r' \)) in both modulus and phase. Even though the quantity \( D(r', \rho') \) is convolved with \( G_d(r', \rho') \) which for this value of \( \rho' \) is the shape of one circle occluded with another (see Fig. 1b), we need hardly perform a deconvolution of \( G(r', \rho') \) in

Fig. 3. Incoherent CBED pattern taken from graphite.

Fig. 4. Coherent microdiffraction pattern taken from graphite. Note the fringes in the disc overlap regions.

Fig. 5. Slice (at constant \( \rho' \), plotted as a function of \( r' \)) in both modulus and phase.
a few delta functions. $G(r', p')$ has now given us a measure, via Eq. (3), of all pairs of phase differences between the diffracted orders which we can use to phase each diffracted order [14]. In this example, we have information corresponding to the first four reflections of graphite, implying

---

**Fig. 5.** Plot of $G(r', p')$ as a function of $r'$ for $|p'| = (0.34 \, \text{nm})^{-1}$ in: (a) magnitude; (b) phase, the greyscale spans the range $-\pi$ (black) to $+\pi$ (white).

**Fig. 6.** Reconstructed exit wave using only the first-order reflections from graphite in: (a) magnitude; (b) phase.
an effective resolution of 0.17 nm. It should be remembered that all this is being transmitted through a machine with a conventional bright-field resolution of 0.42 nm and a conventional coherence envelope of about 0.34 nm. In fact, the meaning of the reconstruction we obtain by synthesizing such a large aperture from this data is not entirely straightforward. For completeness, we show the exit-wave reconstruction we obtain by using only the first-order reflections (Fig. 6: in one dimension in phase and magnitude), and that obtained after adding in the Fourier components from the second-order reflections (Fig. 7). We note that in the latter, the phase component has become stronger and the magnitude weaker, and the phase could be taken as an estimate of the projected atomic potential at 0.17 nm resolution. However, such direct interpretation is possibly dangerous, because we have not accounted for propagation effects throughout the depth of the specimen. We defer a detailed analysis of how these effects complicate the direct interpretation; in the present paper we are concerned with re-defining the information limit in this scattering geometry.

The Fourier transform of a weak phase object contains a very-high-amplitude zero-order peak, and a much weaker scattered part. From Eq. (3) it can be deduced that $D(r', \rho')$ has high magnitude on $r' = 0$ and $\rho' = r'$. Remembering that $G(r', \rho')$ is a convolution with respect to $r'$ between $G_J(r', \rho')$ and $D(r', \rho')$. Fig. 2 shows the amplitude of $G(r', \rho')$ for parallel components of $\rho'$ and $r'$. Note that there is no information for $|\rho'| > 2\alpha$. For a weak phase object the plane defined by $\rho' = 2r'$ contains the Fourier transform of the object function in amplitude and phase with $\frac{3}{2}\alpha < |\rho'| < 2\alpha$, that is up to a resolution of twice the aperture radius [6], which corresponds to a very easy, computationally cheap super-resolution method, especially since no knowledge is needed of defocus, spherical aberration or astigmatism. For $|\rho'| < \frac{1}{2}\alpha$, interference from the opposite side-band is included so information from other values of $r'$ has to be used, such as by deconvolving $G_J(r', \rho')$ from $G(r', \rho')$.  

Fig. 7. Reconstructed exit wave using both the first- and second-order reflections from graphite in: (a) magnitude; (b) phase.
3. Chromatic aberrations

Let us first consider the effect of various components of partial coherence in this data set starting with chromatic aberrations which result from the thermal spread of electrons emitted by the tip or filament, instabilities in the accelerating high-tension supply and instabilities in the current supply to the objective lens. The effect is a spread of defocus values which must be incoherently summed to give the resulting image. The incoherent defocus spread $\Delta$ is given by

$$\Delta = C \left[ \frac{\sigma^2(V_0)}{V_0^2} + \frac{4\sigma^2(I_0)}{I_0^2} + \frac{\sigma^2(E_0)}{E_0^2} \right]^{1/2},$$

(6)

where $\sigma^2(V_0)$, $\sigma^2(I_0)$ and $\sigma^2(E_0)$ are the variances of the accelerating voltage, $V_0$, objective lens current, $I_0$, and emitted electron energies, $E_0$, respectively.

We integrate $|M(r', \rho)|^2$ over $z$, which is equivalent to integrating $G(r', \rho')$ over $z$. Since the only part of $G(r', \rho')$ which has a dependence on $z$ is $G_A(r', \rho')$, it is simplest to consider the effect of chromatic aberrations on $G_A(r', \rho')$. Integrating $G_A(r', \rho')$ over $z$ and ignoring any physical apertures we have

$$G_{hr}^A(r', \rho') = \int \exp\left[i \chi(r', z - z_1) - i \chi(r' + \rho', z - z_1)\right] f(z_1) \, dz_1,$$

(7)

where the defocus fluctuation is represented by the normalised function $f(z)$, and $z_1$ is the dummy variable of integration. The phase of the aperture function, $\chi$, has been written here with $z$ as an explicit argument. A physically reasonable form for $f(z)$ is

$$f(z) = \frac{1}{\Delta \sqrt{2\pi}} \exp\left(-\frac{z^2}{2\Delta^2}\right).$$

(8)

Since $\chi$ has a linear dependence on $z$, $G_{hr}^A(r', \rho')$

$$= G_A(r', \rho') \int f(z_1) \times \exp\left[i z_1 \left( \frac{\partial \chi}{\partial z} \bigg|_{r' + \rho', z} - \frac{\partial \chi}{\partial z} \bigg|_{r', z} \right)\right] \, dz_1,$$

(9)

and we can identify the integral as being the Fourier transform of $f(z)$, thus

$$G_{hr}^A(r', \rho') = G_A(r', \rho') \times F \left[ \frac{1}{2\pi} \left( \frac{\partial \chi}{\partial z} \bigg|_{r' + \rho', z} - \frac{\partial \chi}{\partial z} \bigg|_{r', z} \right) \right],$$

(10)

substituting for $\chi$ using Eq. (5) and $F$ using the Fourier transform of Eq. (8) gives

$$G_{hr}^A(r', \rho') = G_A(r', \rho') \times \exp\left(-\frac{1}{2} \pi^2 \lambda^2 \Delta^2 \left[ 2r' \cdot \rho' + |\rho'|^2 \right] \right).$$

(11)

We can see that the effect of chromatic defocus spread is to modulate $G_A(r', \rho')$ by a four-dimensional envelope function. Not surprisingly the form of the envelope is the same as that derived for CTEM tilted beam imaging of weak scattering objects [15], and also that of the transmission cross-coefficient for crystalline objects [16]. The envelope can be seen to have no dependence on defocus. Fig. 8 shows the temporal coherence envelope function calculated for $\Delta = 10$ nm, $\lambda = 3.7$ pm. The defocus spread chosen is rather pessimistic for the STEM instrument used in these experiments, but conveniently illustrates the form of the envelope. For comparison with Fig. 1 the same planes in the four-dimensional function are shown. Fig. 8c shows a circle where no attenuation of $G_A(r', \rho')$ occurs. In CTEM tilted illumination imaging, spatial frequencies lying on this “achromatic circle” are preferentially transferred [17]. For a given value of $r'$ it can be seen that at $\rho' = -2r'$, $G_A(r', \rho')$ suffers no attenuation since this point is on the achromatic circle, even in the presence of astigmatism; Fig. 8a shows this effect clearly. The $\rho' = 2r'$ reconstruction scheme, mentioned earlier, makes use of only this plane in $G_A(r', \rho')$ and is therefore completely insensitive to chromatic defocus spread. For a crystalline specimen using the centre of overlap between two diffracted orders to determine their phase difference is similarly insensitive because it has no dependence on defocus [13].

The robustness of the algorithm to chromatic aberrations can be understood by realising that
using the $p' = 2r'$ reconstruction scheme uses only the interference between beams which have passed through diametrically opposite points in the objective lens, and have therefore experienced the same phase change.

4. Partial spatial coherence due to a finite detector pixel size

The effect of a finite detector pixel size in the microdiffraction plane is a form of partial spatial coherence. It is convenient to think of it as perfectly incoherent illuminating beam divergence in the equivalent CTEM experiment [18]. It can be modelled as a convolution, with respect to the $r'$ vector, of $|M(r', p')|^2$ with a normalised function $P(r')$ representing the pixel shape. Here we will only consider the first-order effects of the pixel size, the second-order terms for axial illumination in CTEM are discussed in Ref. [19] and are assumed to be negligible here. The next stage of the phase-retrieval/super-resolution algorithm is to take the Fourier transform with respect to $p'$, so the convolution mentioned above is unaffected. Thus

$$G_{\text{det}}^\text{det}(r', p') = G(r', p') \otimes_r P(r') = G_\lambda(r', p') \otimes_r D(r', p') \otimes_r P(r').$$

(12)

In theory $P(r')$ could be deconvolved to give a coherent data set. It is more interesting, though, to observe how information is lost in $G(r', p')$. Since the convolution operator is associative we can write the convolution explicitly, ignoring any physical apertures, as

$$G_{\text{det}}^\text{det}(r', p') = D(r', p') \otimes_r \int \exp\{i\chi(r' - a') - i\chi(r' - a' + p')\} P(a') \, da'.$$

(13)

where $a'$ is a dummy variable of integration. We can now expand to first order

$$G_{\text{det}}^\text{det}(r', p') = D(r', p') \otimes_r \left\{ G_\lambda(r', p') \int P(a') \right\} \exp\left\{ i\chi \left[ \frac{\partial \chi}{\partial q'} r' + p' - \frac{\partial \chi}{\partial q'} r' + p' \right] \right\} \, da'$$

(14)
Fig. 9. Detector coherence envelope for a square detector pixel with side 1 mrad, $\lambda = 3.7 \text{ pm}$, $C_r = 3.1 \text{ mm}$: (a–c) for $z = 0 \text{ nm}$, $z = -128.5 \text{ nm}$ (Scherzer defocus) and $z = -300 \text{ nm}$, respectively, along corresponding parallel components of $r'$ and $\rho'$, the other components of both $r'$ and $\rho'$ are zero; (d–f) for $z = 0$, $-128.5$ and $-300 \text{ nm}$, respectively, as a function of $r'$ for $\rho_1 = 3 \text{ nm}^{-1}$; (g–i) for $z = 0$, $-128.5$ and $-300 \text{ nm}$, respectively, as a function of $\rho'$ for $r_1 = 2 \text{ nm}^{-1}$.
in which the integral is simply the Fourier transform of $P(r')$, thus

\[ G_{\text{det}}(r', \rho') = D(r', \rho') \otimes_r \left( G_\delta(r', \rho') \right) \times p \left( \frac{1}{2\pi} \left[ \frac{\partial X}{\partial q} \mid_{r+r'} - \frac{\partial X}{\partial q} \mid_{r} \right] \right). \] (15)

We can see that, to first order, the effect of finite detector size is to modulate $G_{\delta}(r', \rho')$ by a four-dimensional envelope function. We can think of this envelope function as being the set of CTEM illumination divergence coherence envelopes, for every illumination tilt $r'$. Fig. 9 shows the envelopes for Gaussian focus, Scherzer defocus and a large amount of defocus. For comparison with Fig. 1 the same planes in the four-dimensional function are shown. Fig. 9h shows how the envelope can act as a “virtual aperture” [20] shifted by the degree of tilt, which is represented by $r'$. The virtual aperture is more appropriate in tilted illumination experiments because physical ones tend to charge up [21].

It is important to note that the envelope is dependent on defocus. This dependency can be used to detect and determine the CTEM illumination divergence coherence envelopes [22] For an illumination tilt of $r'$, using a defocus of

\[ z = -C_x \lambda^3 |r'|^2 \] (16)

passes the point $\rho' = -2r'$ with no attenuation; a criterion used by some workers (reviewed by Wade [23]) in tilted illumination CTEM imaging because the point is also unaffected by chromatic defocus spread, being on the achromatic circle.

The phase-retrieval/super-resolution algorithm uses information across a range of $\rho'$ values. The criterion for focus is thus one which minimises

\[ \left[ \frac{\partial X}{\partial q} \mid_{r+r'} - \frac{\partial X}{\partial q} \mid_{r} \right] \]

over as large a range as possible. A similar argument applies as a focusing criterion in off-axis holography [24]. Observation of Fig. 9 shows that Scherzer defocus gives a larger coherence envelope than Gaussian focus. However, at higher values of defocus “holes” appear in the envelope. Though certain higher values of $\rho'$ are strongly transferred, which can be useful for a crystalline specimen. Scherzer defocus gives a virtual aperture at a radius of approximately 3 mm for $\lambda = 3.7$ pm and $C_x = 3.1$ mm, and the magnitude of $G_{\delta}(r', \rho')$ is very similar to that for a physical objective aperture as shown in Fig. 1c. There is no longer a need for a physical objective aperture. We can simply collect shadow images as a function of probe position and process them as described in Section 2 with the microscope at Scherzer defocus.

The existence of the virtual aperture has been verified experimentally using a STEM. In the absence of an objective aperture a shadow image is observed in the far-field [25–27]. Shadow images of graphite were recorded as a function of probe position $\rho$, a typical shadow image at approximately Scherzer defocus is shown in Fig. 10. A Fourier transform was taken of the resulting data set with respect to $\rho$ to get $G(r', \rho')$. As in the experiment described earlier, which used a physical aperture, amplitude is only observed in

Fig. 10. Typical shadow image of a crystal of graphite taken at approximately Scherzer defocus.
Fig. 11. Plot of $G(r', \rho')$ for data collected without an objective aperture, as a function of $r'$ with $|\rho'| = (0.34 \text{ nm})^{-1}$, in: (a) magnitude; (b) phase, the greyscale spans the range $-\pi$ (black) to $+\pi$ (white).

$G(r', \rho')$ at a value of $\rho'$ corresponding to the first-order lattice spacing. The magnitude and phase of $G(r', \rho')$ at this value of $\rho'$ is shown in Fig. 11 as a function of $r'$. The interference between the first and second order can be seen to be separate from that between the zero- and first-order diffracted beams, the virtual aperture limiting the angular range of the interference.

Fig. 12. Magnitude of $G(r', \rho')$ for data collected without an objective aperture and with a heavily defocused probe, as a function of $r'$ with $|\rho'| = (0.34 \text{ nm})^{-1}$. The weak fringes observed are caused by three-beam interference.

Fig. 12 shows the magnitude of $G(r', \rho')$ for a highly defocused probe. It can be seen that the interference region now has a hole as predicted above. Weak fringes can be observed within the hole, these are caused by three-beam interference and can be thought of as part of a four-dimensional contrast transfer function [6].

We can understand the virtual aperture physically as follows: beyond a certain angle in the shadow image the size of the interference fringes become smaller than the detector pixel size and so the change as the probe is scanned is not detected. The dependence of the magnification in the shadow image on the phase of the aperture function has been discussed by Cowley [25].

5. Partial spatial coherence due to a finite electron source

The theory described in Section 2 above assumed an infinitesimal electron source. In reality
this can never be the case for an electron source of finite brightness. Rodenburg and Bates (Eq.
(30) in Ref. [3]) have already considered this situation, but we will review it here for complete-
ness. A finite electron source in STEM is equivalent to a finite detector pixel in the image plane
of a CTEM [18]. In the notation used above we can model this as a convolution of the $|M(r', p')|^2$
data set with respect to the $p$ variable:

$$|M^\text{cont}(r', p)|^2 = |M(r', p)|^2 \otimes s(p), \quad (18)$$

where $s(p)$ is a normalised function describing the effective source intensity as measured at the
specimen, that is taking into account the demagnification effects of the objective and pre-objective
lenses. As described earlier, the super-resolution/phase-retrieval algorithm goes on to take the
Fourier transform of this data set with respect to the $p$ vector. Eq. (18) is transformed into

$$G^\text{cont}(r', p') = G(r', p') \ast S(p'). \quad (19)$$

where $S(p')$ is the Fourier transform of $s(p)$.

Eq. (19) implies that the effect of finite source size in STEM is to modulate $G(r', p')$ by an
envelope function, $S(p')$, in just the $p'$ direction.

If we were possible to deconvolve $G_\alpha(r', p')$ from $G^\text{cont}(r', p')$ to leave just $D^\text{cont}(r', p')$ with suffi-
cient accuracy, we could determine the magni-
tude of $S(p')$ using

$$|S(p')|^2 = \frac{|D^\text{cont}(r', p')|^2}{D^\text{cont}(r', 0) D^\text{cont}(r', p', 0)}. \quad (20)$$

If we assume that the envelope $S(p')$ is real and positive for all $p'$ (i.e. that $s(p)$ is symmetric and
smoothly varying) then Eq. (20) completely describes $S(p')$. We can simply divide Eq. (19) by
$S(p')$, thereby deconvolving the effects of finite source size completely. However, the accuracy of
Eq. (20) is entirely dependent on the accuracy of the deconvolution of $G_\alpha(r', p')$ from $G(r', p')$,
which is not very easy in practice.

If the envelope described by $S(p')$ destroys informa-
tion within the finite detector size envelope then it will limit the resolution of the $p' = 2r'$/
scheme. At Scherzer defocus this situation occurs
when the effective source size is larger than about 0.2
nm for $\lambda = 3.7$ nm and $c_1 = 3.1$ mm. It does not
limit ultimate resolution if “stepping out” in the
$r'$ direction is used. In the case of a perfect thin
egin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig13.png}
\caption{Plot of $G^\text{cont}(r', p')$ as a function of $r'$ with $|p'| = 0.34$ nm$^{-1}$ and a simulated effective source diameter of 0.25 nm, in: (a) magnitude; (b) phase, the greyscale spans the range $-\pi$ (black) to $+\pi$ (white). Compare with Fig. 5.}
\end{figure}
crystal as long as the effective source size is smaller than first-order lattice spacing then the ultimate resolution is not limited. Consider the experiment on graphite discussed in Section 2. We have simulated an incoherent probe by convolving the measured intensity data set, $|M(r', \mathbf{p})|^2$, with an effective source with a width of 0.25 nm which is more than half the first-order lattice spacing. The magnitude and phase of the corresponding $G^{\text{conv}}(r', \mathbf{p'})$ is shown in Fig. 13. It can be seen that all the information is retained, but with a reduced magnitude compared to Fig. 5a. The diffracted orders can still be phased to give a resolution of 0.17 nm, which is smaller than the effective source size. We can understand this condition physically as follows: consider the microdiffraction pattern of a perfect thin crystal. When the probe is moved by one lattice spacing, the fringes in the disc overlap are all simultaneously shifted by one fringe spacing in a direction parallel to the probe movement. The effect of a finite source size is equivalent to adding the intensity of all the microdiffraction patterns for probe positions within the effective source size. It is now obvious only when the effective source size is comparable to the lattice spacing is the coherent interference information lost.

6. Conclusions

We have described the coherence envelope which affects the ability to perform super-resolution aperture synthesis in the STEM mode. There are three important quantities: the effective source size, having taken into account demagnification factors in the probe-forming optics; the effective detector element size, having taken into account any post-specimen optics; and the maximum gradient in the phase of the transfer function of the probe-forming lens. The coherence envelope can be made completely insensitive to chromatic and lens instability effects.

In a particular microscope, the crucial quantity is the width (in reciprocal units) of any single "step out" in the aperture synthesis stage of the reconstruction algorithm. For the purposes of ease of interpretation, it is convenient that this quantity should be as wide as possible, so as to allow for super-resolution projection imaging [8,9]. In Table 1, we tabulate the size of this step width for various instrument parameters with the square detector pixel of side 1 mrad used for the earlier calculations. An obvious method for increasing the width of the "step out" would be to reduce the detector pixel size. However, the width of the "step out" varies as only the cube-root of the pixel size; and, naturally, the smaller the detector element size, the much larger is the requirement for the computing and storage needed for the inversion algorithm. However, since high-resolution imaging in this mode need not rely on extremely small $C_s$, it may be appropriate to worsen the lens parameters (so as to improve the specimen tilt facilities), or lower the accelerating voltage so as to reduce the cost of the machine, yet compensate for all these effects by enlarging the computing facilities attached to the microscope. Since major computing hardware is now so cheap, this latter strategy may be the way ahead for electron imaging.

**Acknowledgements**

The authors are grateful to the Science and Engineering Research Council for financially supporting this work. J.M.R. is also grateful to the Royal Society (London) for financial support. The authors would also like to thank K. Ishizuka for his helpful and critical refereeing of this paper.

<table>
<thead>
<tr>
<th>$V_0$ (kV)</th>
<th>$C_s$ = 3.1 mm</th>
<th>$C_s$ = 1.0 mm</th>
<th>$C_s$ = 0.5 mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.20 nm</td>
<td>0.14 nm</td>
<td>0.11 nm</td>
</tr>
<tr>
<td>300</td>
<td>0.12 nm</td>
<td>0.09 nm</td>
<td>0.07 nm</td>
</tr>
<tr>
<td>1 MV</td>
<td>0.07 nm</td>
<td>0.05 nm</td>
<td>0.04 nm</td>
</tr>
</tbody>
</table>
References