Ultramicroscopy 151 (2015) 37–45

Contents lists available at ScienceDirect

Ultramicroscopy

journal homepage: www.elsevier.com/locate/ultramic

Optimising electron holography in the presence of partial coherence and instrument instabilities

Shery L.Y. Chang *, Christian Dwyer *, Chris B. Boothroyd, Rafal E. Dunin-Borkowski

Ernst Ruska-Centre for Microscopy and Spectroscopy with Electrons and Peter Grünberg Institute, Forschungszentrum Jülich, Jülich 52425, Germany

ARTICLE INFO

Article history:
Received 4 July 2014
Received in revised form
10 November 2014
Accepted 10 November 2014
Available online 25 November 2014

Keywords:
Electron holography
Phase error
Spatial coherence
Instrument instability

ABSTRACT

Off-axis electron holography provides a direct means of retrieving the phase of the wavefield in a transmission electron microscope, enabling measurement of electric and magnetic fields at length scales from microns to nanometers. To maximise the accuracy of the technique, it is important to acquire holograms using experimental conditions that optimise the phase resolution for a given spatial resolution. These conditions are determined by a number of competing parameters, especially the spatial coherence and the instrument instabilities. Here, we describe a simple, yet accurate, model for predicting the dose rate and exposure time that give the best phase resolution in a single hologram. Experimental studies were undertaken to verify the model of spatial coherence and instrument instabilities that are required for the optimisation. The model is applicable to electron holography in both standard mode and Lorentz mode, and it is relatively simple to apply.

1. Introduction

Off-axis holography in the transmission electron microscope (TEM) is an established technique for measuring the electrostatic and magnetic properties of materials and devices. The technique reconstructs the phase shifts experienced by the electron wavefield and uses them to map the spatially varying electric or magnetic field. In the pursuit of measuring the increasingly weaker electric and magnetic fields generated from nanomaterials, the necessary improvements in the resolution of the reconstructed phase have been pursued using various strategies, which can be loosely categorised according to instrumental improvements, and improvements in data acquisition and processing.

There exists a fairly extensive body of literature reporting phase resolution improvements within both of the above-mentioned categories. In general, the phase resolution can be improved only by increasing the *coherent* electron dose. On the instrumentation side, higher coherent doses have been achieved by brighter electron sources [1,2] and the use of elliptical illumination [3,4]. Improvements in microscope stability enable larger doses via longer exposure times [5]. The improvement of a charge-coupled device (CCD) camera's modulation transfer function (MTF) can also increase the detectable coherent dose [6]. On the data acquisition and processing side, a greater dose and hence better phase resolution has been achieved by the use of multiple holograms [7–9], which applies also to the case of phase-shifting holography [10–12].

Considering the number of available methods for improving the phase resolution, it is important to understand the dominant factors that limit the phase resolution. Typically, many of the experimental parameters are pre-determined by the requirements of the specimen. These include the hologram fringe spacing (which determines the spatial resolution), overlap width (which determines the field of view), and magnification (which should be as high as the overlap width allows). These parameters are therefore regarded as essentially fixed. The remaining parameters with which we can optimise phase resolution can be grouped into two categories, namely the partial spatial coherence and the instrument instabilities. These parameters are controlled via the electron dose and exposure time, respectively.

In this paper, we describe a simple, yet accurate, model capable of predicting the dose rate and exposure time that give the best phase resolution in a single hologram. To make the presentation tractable, we have restricted our attention to the instrumental factors affecting holography, and have not concerned ourselves explicitly with factors associated with the specimen. Hence our results reflect the best-possible phase resolution that can be achieved under given conditions on a particular instrument. In the presence of specimen drift and/or dose-dependent specimen damage, the optimum dose rate given here remains entirely valid, while the

* Corresponding authors.
E-mail addresses: shery.chang@fz-juelich.de (S.L.Y. Chang), c.dwyer@fz-juelich.de (C. Dwyer).

http://dx.doi.org/10.1016/j.ultramic.2014.11.019
0304-3991/

optimum exposure time may need to be reduced accordingly.

In light of the considerable body of literature describing the theories and experimental factors governing the phase resolution in off-axis electron holography [4,7,13–17], our work requires some justification: our aim is to provide a simple and practical methodology, as free as possible of unnecessary details. In particular, the models for spatial coherence and instrument instabilities are kept as simple as possible. Furthermore, we describe a minimal experimental dataset that can be used to predict the optimum conditions for all combinations of fringe spacings, overlap widths and magnifications. Our results can also be applied, with minimal modification, to the case where multiple holograms are used.

This paper is organised as follows: Section 2 provides some background on the concept of a phase error in electron holography. In Section 3 we outline the theoretical model used for predicting the optimum conditions. Section 4 describes our experimental setup and processing methods. Our results and discussion are presented in Section 5. In Section 6 we discuss the extension to elliptical illumination before concluding in Section 7.

2. The phase error

The phase resolution, herein referred to as the phase *error*, determines the minimum difference that can be distinguished in the reconstructed phase (here we are concerned with statistical errors rather than systematic ones). Fig. 1(a) illustrates the phase error associated with an arbitrary point in the reconstructed wave function. In the ideal case, each point of the wave function would correspond to a point in the Argand plane. However, due to the finite electron dose (among other reasons), there is always a statistical error associated with the complex value ψ. In Fig. 1(a) this error is represented as a cluster of points spread symmetrically around the nominal value ψ, the points corresponding to the values obtained by repeated independent measurements. The phase error is typically defined as the standard deviation $\Delta\phi$ of the repeated phase measurements [13–15].

For the case illustrated in Fig. 1(a), the phase error is given to good approximation by $\Delta\phi \approx \Delta\psi/A_{mean}$. For very low doses, however, the noise in the reconstructed phase grows to the extent that the phase error represented by this simple formula becomes ill-defined (in extreme cases the phase error so calculated will exceed 2π). This problem can be remedied by using the standard deviation associated with the cosine and sine of the phase, as illustrated in Fig. 1(b). The latter definition has the benefit of remaining well defined for arbitrarily low doses. The two definitions are equivalent for sufficiently high doses.

3. Theory

The interference pattern produced by two partially coherent plane waves $e^{2\pi i\mathbf{k}_1\cdot\mathbf{x}}/\sqrt{2}$ and $e^{2\pi i\mathbf{k}_2\cdot\mathbf{x}}/\sqrt{2}$, as measured by a pixelated electron detector in an off-axis holographic setup, is described by the expression

$$N(\mathbf{x}) = CN_e\left[1 + \frac{V}{2}e^{2\pi i\mathbf{k}_1\cdot\mathbf{x}}e^{-2\pi i\mathbf{k}_2\cdot\mathbf{x}} + \frac{V^*}{2}e^{-2\pi i\mathbf{k}_1\cdot\mathbf{x}}e^{2\pi i\mathbf{k}_2\cdot\mathbf{x}}\right]$$
$$= CN_e\left[1 + |V|\cos(2\pi(\mathbf{k}_1 - \mathbf{k}_2)\cdot\mathbf{x} + \arg V)\right]. \quad (1)$$

This expression describes a set of cosinusoidal fringes sitting on a constant background, expressed in terms of detector signal $N(\mathbf{x})$, where N_e is the average number of detected electrons per pixel, C is a constant which we describe below, the complex number V weights the interference terms and obeys $0 \le |V| \le 1$, and $\arg V$ denotes the complex argument (or phase) of V. The number V incorporates any factors that lead to a damping of the interference fringes in a relative sense, which include the partial spatial coherence of the beam, instabilities of the instrument, and the less-than-perfect modulation transfer function (MTF) of the detector. The magnitude $|V|$ is the visibility of the interference fringes (also termed the fringe contrast). The constant C equals the average signal output from the detector per incident electron, and so it incorporates factors such as the less-than-perfect detector quantum efficiency (DQE) and the detector gain G.

3.1. Phase error and the effective signal

Following Fourier processing of the off-axis hologram, the statistical phase error in a given pixel in the reconstructed wave function is given approximately by the expression

$$\Delta\phi \approx \left(\frac{2G}{N_{\text{eff}}}\right)^{1/2}, \quad (2)$$

where N_{eff} is the *effective* signal per pixel [13,14]. The latter is defined as

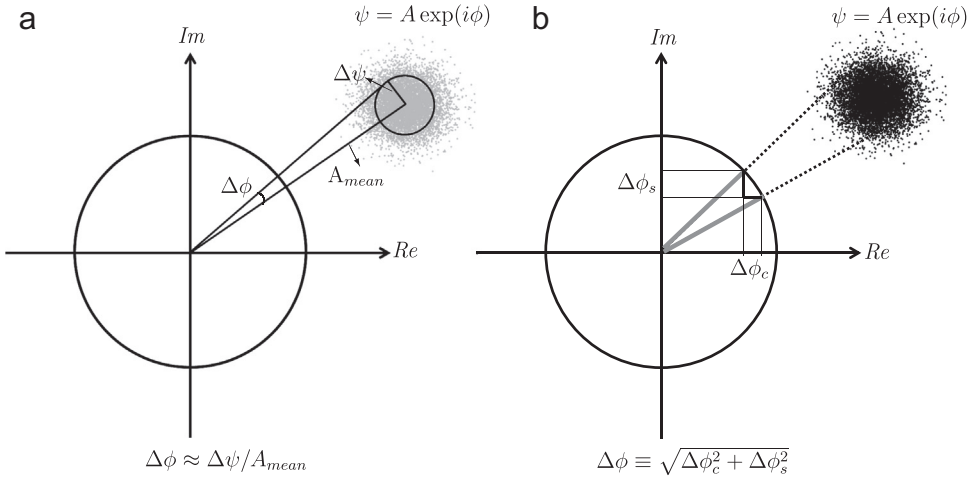

Fig. 1. Schematic representations of the phase error associated with a given point in the reconstructed wave function ψ. (a) Conventional definition of the phase error, applicable when the dose level is sufficiently high. (b) Definition of the phase error in terms of the cosine and sine of the phase. The definition in (b) remains well-defined for arbitrarily low doses. For high doses these definitions become equivalent.

$$N_{\text{eff}} = |V|^2\langle N(\mathbf{x})\rangle. \tag{3}$$

The phase error in Eq. (2) is understood as being inversely proportional to the Poisson noise of N_{eff}/G, which is the effective number of detected electrons per pixel. Eq. (2) holds provided that $N_{\text{eff}}/G \gg 1$. If $N_{\text{eff}}/G \lesssim 1$, then Eq. (2) diverges and we should employ the more general definition of phase error outlined in Fig. 1(b).[1]

The effective signal N_{eff} is a crucial concept in the present work and the appearance of $|V|^2$ in the definition of N_{eff} warrants a remark. Although this expression for the effective signal has been derived previously [14], intuitively it might be expected that the effective signal should be proportional to the amplitude of the interference fringes, which would make N_{eff} proportional to $|V|$, and not $|V|^2$. However, in addition to the Poisson noise associated with the interference fringes, it is crucial to consider the Poisson noise contributed by the constant background. Consideration of these contributions shows that the effective signal is proportional to $|V|^2$.

From Eq. (2), it is easily understood that the optimum condition for holographic phase measurements corresponds to maximising N_{eff}. In order to proceed, we must adopt a model to describe the visibility $|V|$. In accordance with the aims of the present work, the model used here is kept as simple as possible while still affording sufficient accuracy:

$$|V| = V_0 V(t) S(\hat{N}). \tag{4}$$

where V_0 is the time- and dose-independent part of the visibility, which arises from factors such as the detector MTF, $V(t)$ is the time-dependent part of the visibility, which arises from instabilities that are slower than the fastest exposure time, and $S(\hat{N})$ is the spatial coherence envelope of the wavefield, which is written here as a function of the average signal per pixel per unit time: $\hat{N} = N/t$. Note that Eq. (4) is separable in $\hat{N}$ and t, which allows $\hat{N}$ and t to be optimised independently.

3.2. Model of partial spatial coherence

The partial spatial coherence of the wavefield arises primarily from finite source effects and instabilities in the gun and illumination systems. Together, these effects can be grouped into an effective incoherent source distribution. For holography, we employ Köhler-like illumination, so that the relevant spatial coherence envelope is, according to the Van Cittert–Zernike theorem [18], given by the Fourier transform of the effective source. Assuming round illumination, we adopt the following phenomenological model for the partial spatial coherence envelope:

$$S(\hat{N}) = e^{-\alpha(w/\delta)\hat{N}^{1/2} - \beta(w/\delta)^2\hat{N}}, \tag{5}$$

where α and β are constants to be determined, w is the interference width, and δ is the effective pixel size as determined by the magnification (not to be confused with the physical pixel size in the camera). This spatial coherence model corresponds to a rotationally symmetric effective source having the form of a bivariate Cauchy distribution (the first term in the exponent) convoluted with a Gaussian distribution (the second term in the exponent).

[1] We also note that, strictly, the expression in Eq. (2) applies only if (1) the reconstruction uses a mask that is localised in real space, and (2) the detector preserves the Poisson statistics of the electron beam [16]. While these conditions are not strictly satisfied in our experiments, they are satisfied sufficiently well that meaningful and accurate results are obtained, as will be verified by the results presented in later sections.

The commonly used Gaussian coherence model [17], although especially simple, did not produce a satisfactory fit to our experimental data. This observation is reminiscent of source size measurements in scanning transmission electron microscopy, where departures from a Gaussian source shape have been observed [19–21]. The model in Eq. (5) is similar to that used in Ref. [21], but it contains only two parameters (α and β) instead of three. While the model of Ref. [21] is also capable of providing a good fit to our experimental data, the simpler form of our model has advantages in mathematical manipulations.

3.3. Model of time-dependent visibility

The primary instabilities contributing to $V(t)$ are likely to be the biprism position and/or beam tilt, which produce a movement of the holographic fringes during the exposure time. Here we make no attempt to analyse the origins of the fringe movement, but seek only to model it. It is worth noting that movement of the biprism cannot be readily distinguished from beam tilt, and that the effect of either instability is independent of the fringe spacing.

Our analysis of the hologram stability (presented in Section 5.2) reveals that the fringes exhibit both a stochastic motion and a longer-term near-linear drift. For common exposure times, we find that the stochastic motion dominates. Hence the fringe motion can be described by a probability distribution $\rho(x, t)$, which gives the probability of a displacement x at time t. A mathematical analysis shows that such motion gives rise to the following time-dependent visibility:

$$V(t) = \left|\frac{1}{t}\int_0^t \mathrm{d}t' f(t')\right|, \tag{6}$$

where the function $f(t)$ is the Fourier transform of the probability distribution $\rho(x, t)$ evaluated at the carrier frequency:

$$f(t) = \int \mathrm{d}x\, \rho(x, t) e^{-2\pi i x/d}, \tag{7}$$

where d is the fringe spacing. Note that in Eq. (6), $V(t)$ is defined as the modulus of a complex number, which reflects the loss of fringe contrast due to their movement. (The phase that is omitted by Eq. (6) reflects the average apparent shift of the fringes after an exposure time t, though we do not use this phase in our analysis.)

The stochastic motion can be modelled using Langevin theory, which was originally developed to describe the Brownian motion of particles undergoing random collisions. Adopting such a theory, the displacement of the fringes obeys the following probability distribution:

$$\rho_{\text{stoch}}(x, t) = \frac{1}{\sqrt{4\pi\langle x^2(t)\rangle}} e^{-x^2/2\langle x^2(t)\rangle}, \tag{8}$$

where the variance of the displacement is given by

$$\langle x^2(t)\rangle = 2d_c^2\left(t/t_c + e^{-t/t_c} - 1\right), \tag{9}$$

with t_c and d_c being the average time and average distance between "collisions", respectively (in the present context, a "collision" is simply an event where the fringe velocity changes abruptly). An important aspect of the Langevin theory is that it is applicable to times both short and long compared to the collision time t_c. The motion at these times exhibits the correlational and diffusional behaviours, respectively:

$$\begin{aligned} &t \lesssim t_c\text{: } \sqrt{\langle x(t)\rangle} = d_c t/t_c \\ &t \gg t_c\text{: } \sqrt{\langle x(t)\rangle} = \sqrt{Dt}. \end{aligned} \tag{10}$$

where $D = 2d_c^2/t_c$ is a diffusion coefficient. The function $f(t)$ relevant to the time-dependent visibility becomes

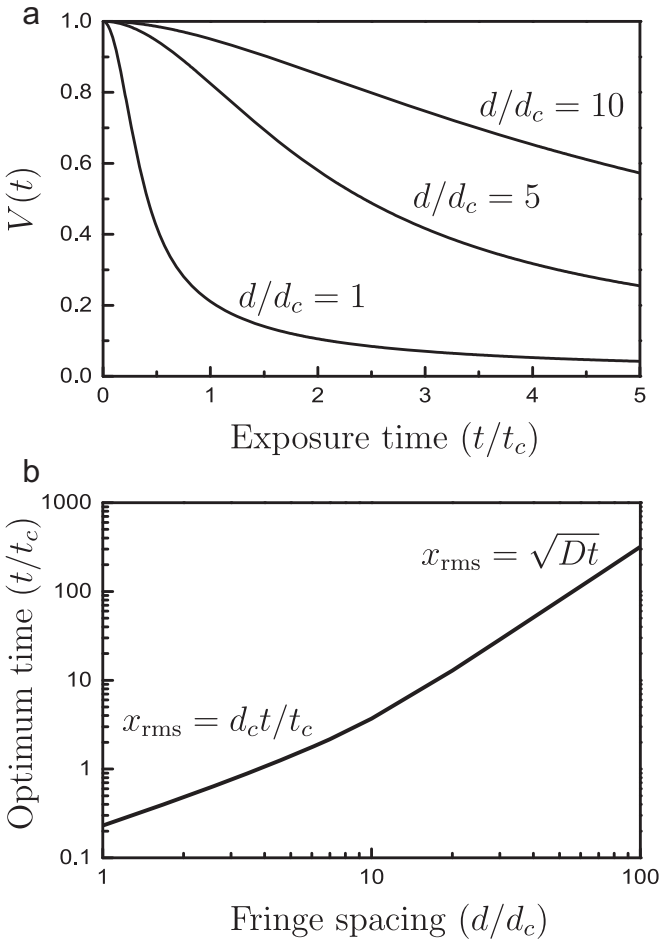

Fig. 2. Theoretical modelling of the time-dependent visibility $V(t)$ due to the stochastic motion of the holographic fringes. (a) $V(t)$ as a function of exposure time for selected fringe spacings, (b) the optimum exposure time as a function of the fringe spacing.

$$f_{\text{stoch}}(t) = e^{-4\pi^2\langle x^2(t)\rangle/d^2} = e^{-(2\pi d_c/d)^2(t/t_c+\exp(-t/t_c)-1)}. \tag{11}$$

Unfortunately, the rather complex dependence on t means that it is not possible to obtain an analytical expression for $V_{\text{stoch}}(t)$ that is applicable at all times. Instead, we have resorted to obtaining $V_{\text{stoch}}(t)$ by numerical integration, with the result shown in Fig. 2a.

The above model for $V(t)$ differs from those of previous works, which employed either a linear displacement model [7] or a diffusion model applicable only to times $t \gg t_c$ [8]. Here we found that the form of the Langevin model, in which t_c separates the regimes of correlational and diffusional motion, was important. It will turn out that commonly used exposure times straddle both regimes.

3.4. Optimum conditions

The experimental parameters that can be adjusted to optimise the conditions for holography consist of the normalised signal $\hat{N}$ (controlled by the gun and illumination lenses), the holographic fringe spacing d (controlled by the biprism voltage), the interference width w (controlled by the biprism voltage and the biprism defocus), the pixel size δ (controlled by the magnification), and the exposure time t. Of these parameters, d, w and δ are dictated by the specimen geometry, which leaves $\hat{N}$, and t as the independent parameters. Hence, given the values of α, β, d_c and t_c applicable to a particular instrument, our goal is to predict, for given values of d, w and δ, the optimum values of $\hat{N}$ and t. These optimum values maximise N_{eff} (or minimise $\Delta\phi$).

A straightforward application of calculus shows that the optimum value of $\hat{N}$ is given by[2]

$$\max_{\hat{N}} N_{\text{eff}}(\hat{N}, t) \Longrightarrow \hat{N} = \frac{\delta^2}{w^2}\left(\frac{\sqrt{\alpha^2+8\beta}-\alpha}{4\beta}\right)^2. \tag{12}$$

A straightforward application of calculus also gives the optimum exposure time:

$$\max_{t} N_{\text{eff}}(\hat{N}, t) \Longrightarrow 2tf_{\text{stoch}}(t) = \int_0^t \mathrm{d}t' f_{\text{stoch}}(t'). \tag{13}$$

While this expression cannot be evaluated analytically, the optimum time can be obtained numerically and is plotted in Fig. 2b.

In practice, the procedure can be made very simple: the parameters α, β, d_c and t_c are determined by fitting data acquired for a specific choice of d, w and δ. Let us denote the latter values by d_0, w_0 and δ_0. The fitting identifies optimal values of $\hat{N}$ and t, which we denote by $\hat{N}_0$ and t_0. Then, for other values of d, w and δ, the optimal $\hat{N}$ is given by

$$\hat{N} = \left(\frac{w_0/\delta_0}{w/\delta}\right)^2 \hat{N}_0, \tag{14}$$

Note that the scaling factor in Eq. (14) does not involve α or β, and so it does not depend on the performance of the instrument (though the actual value $\hat{N}_0$, and hence $\hat{N}$, does). Also note that w_0/δ_0 and w/δ are simply the interference widths measured in units of the effective pixel size. For the optimum t, we must consult Fig. 2b in the general case. However, for fringe spacings less than about $4d_c$, the time-dependent visibility $V(t)$ is limited by correlative motion, and we have the simple and intuitive result $t = (d/d_0)t_0$.

4. Methods

4.1. Experimental set-up

The experiments were carried out using a Titan 80-300 FEG-TEM (FEI Co.) operated at 300 kV. The microscope was equipped with an ultra-bright X-FEG electron gun, two electron biprisms, and a 2k × 2k CCD camera. The upper biprism (not used here) is located in an extra lens inserted between the diffraction lens and the imaging aberration corrector, and the lower biprism is in the selected-area aperture position. Blank holograms (no specimen) were recorded using the lower biprism. The biprism voltage for the primary datasets was set to 150 V, which in *standard* mode (i.e., standard imaging condition with the objective lens on, in contrast to Lorentz mode) corresponds to a fringe spacing of 0.08 nm. The magnification used (450k) corresponds to a pixel size of 0.019 nm per pixel. The overlap width was measured to be 25 nm. The exposure time was 1 s unless specified in the text. All of our experiments employed round illumination, though our methodology can be easily adapted to elliptical illumination (see Section 6).

[2] In passing, we note that this optimum value of $\hat{N}$ corresponds to a degree of spatial coherence given by $S_{\text{opt}} = \exp[\alpha(\alpha - \sqrt{\alpha^2+8\beta})/8\beta - 1/2]$. If we were to assume a Gaussian coherence envelope, then $\alpha = 0$, and we obtain $S_{\text{opt}} = e^{-1/2} \approx 0.61$, in agreement with previous work [17]. We emphasise, however, that a satisfactory fit of our experimental data required nonzero values of both α and β, so that S_{opt} then depends on α and β as above, and hence on the particular instrument. For the instrument used here, the fitted values of α and β give $S_{\text{opt}} \approx 0.45$.

Secondary datasets were obtained for different combinations of the biprism voltage and magnification. Datasets were also taken using Lorentz mode. In the latter mode, the objective lens is turned off, the Lorentz lens is excited, and all other lens settings are maintained. The secondary and Lorentz mode datasets were used only in Section 5.4. The primary dataset was used throughout the paper.

4.2. Processing of holograms

As the processing procedure can affect the measured values of the visibility and the phase error, the following describes our procedure for obtaining the relevant quantities, which include the mean signal (in CCD counts) interference width, fringe visibility, and phase error.

The mean signal was measured from the central regions of the holograms to minimise the influence of the Fresnel fringes. The interference widths, measured relative to the specimen plane, were determined by taking into account both the measured overlap widths in the holograms and the finite width of the biprism wire. The width of the biprism wire relative to the specimen plane was measured by focusing the biprism using the diffraction lens, and then taking into account the associated change in magnification, giving a width of 6.0 ± 0.1 nm.

The fringe visibility and the phase error were measured by the following procedure: a complex wave function having the size of the original hologram was obtained by an inverse Fourier transform of the masked and re-centred sideband. A hard circular mask with a radius of one-third the distance between the centre band and the sideband was employed. Distortions due to the fibre optics of the CCD camera were measured and subsequently removed from the reconstructed phase. Furthermore, a high-pass filter was applied to remove the slowly varying component of the phase arising from effects such as a non-planar wavefront or long-ranged camera distortions. This filter had the Fourier-space representation $\mathrm{HPF}(\mathbf{k}) = 1 + \delta_{\mathbf{k},\mathbf{0}} - e^{-\mathbf{k}^2/2\sigma^2}$, where the width of the Gaussian corresponded to approximately 10 fringe spacings and the delta function has the effect of retaining the wave function's mean value. To minimise the influence of the Fresnel fringes, only the portion of the wave function in the central region of the overlap was considered. The fringe visibility was calculated directly from the average modulus of the wave function. The phase error $\Delta\phi$ was calculated by employing the general definition in Fig. 1(b).

5. Results and discussion

5.1. Partial spatial coherence

The first part of our analysis serves to check how the various lenses in the illumination system affect the spatial coherence of the electron wavefield. From Eq. (4), changes in the spatial coherence are directly related to changes in the fringe visibility. Hence the fringe visibility was measured as a function of excitation of the lenses C1, C2 and Gun Lens, which were varied individually, and all other parameters were kept constant. From Fig. 3(a), it is seen that the measured visibility decreases with increasing signal, and that essentially the same values are measured regardless of which lens is used to control the signal. The fact that the different lenses produce the same result is expected from the point of view of first-order optics, where beam intensity can be gained only at the loss of spatial coherence, in such a way that the beam brightness remains conserved. Our observations here rule out significant higher-order optical effects and instabilities at any of the intermediate images of the source, and thus provide assurance that the illumination lenses are interchangeable in their effect.

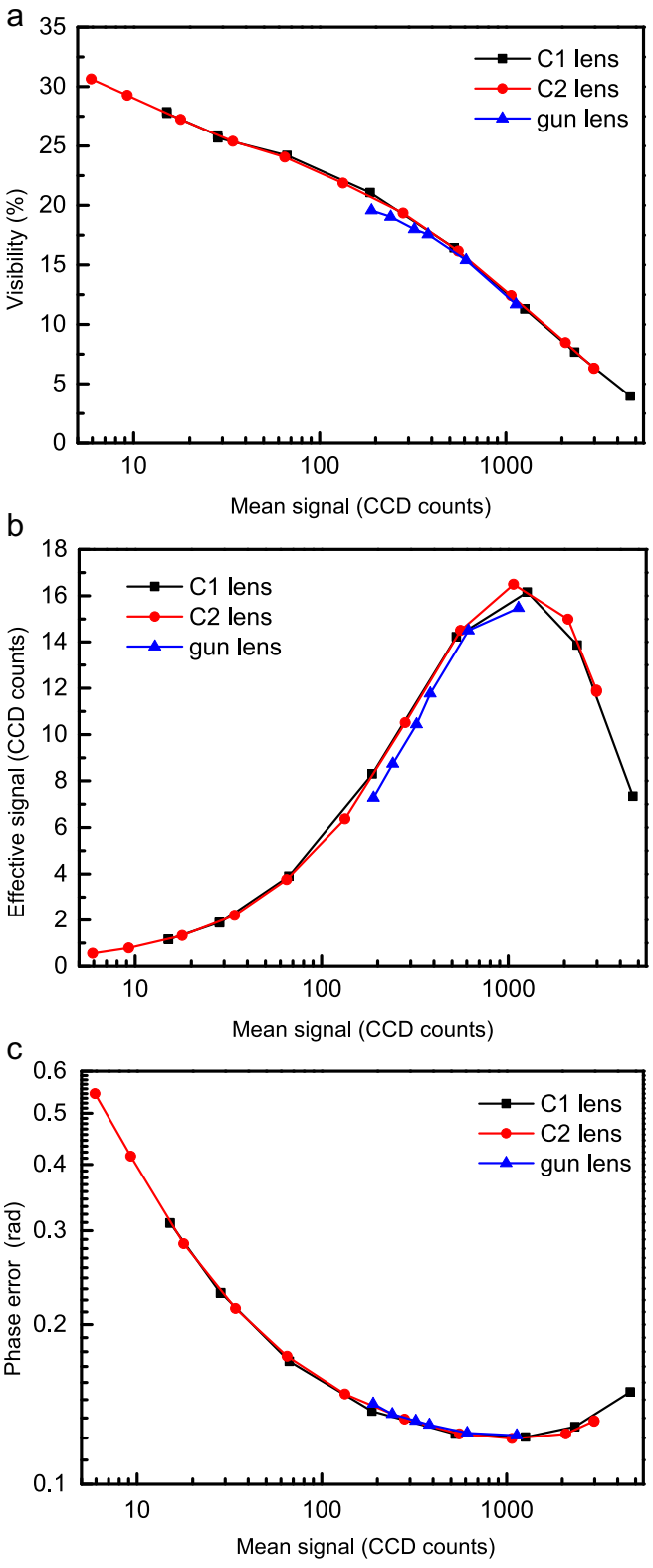

Fig. 3. (a) Fringe visibility, (b) effective signal and (c) phase error as a function of the mean signal, which was controlled by varying the illumination lenses C1, C2 and gun lens. A biprism voltage of 150 V, exposure time of 4 s, and magnification of 450k apply throughout.

Fig. 3(b) and (c) shows the corresponding measurements of the effective signal and the phase error, respectively. The effective signal is given by $N_{\mathrm{eff}} = |V|^2 N$, which contains two competing factors: on the one hand, a stronger signal may result in a larger effective signal. On the other hand, a weaker signal may also lead to

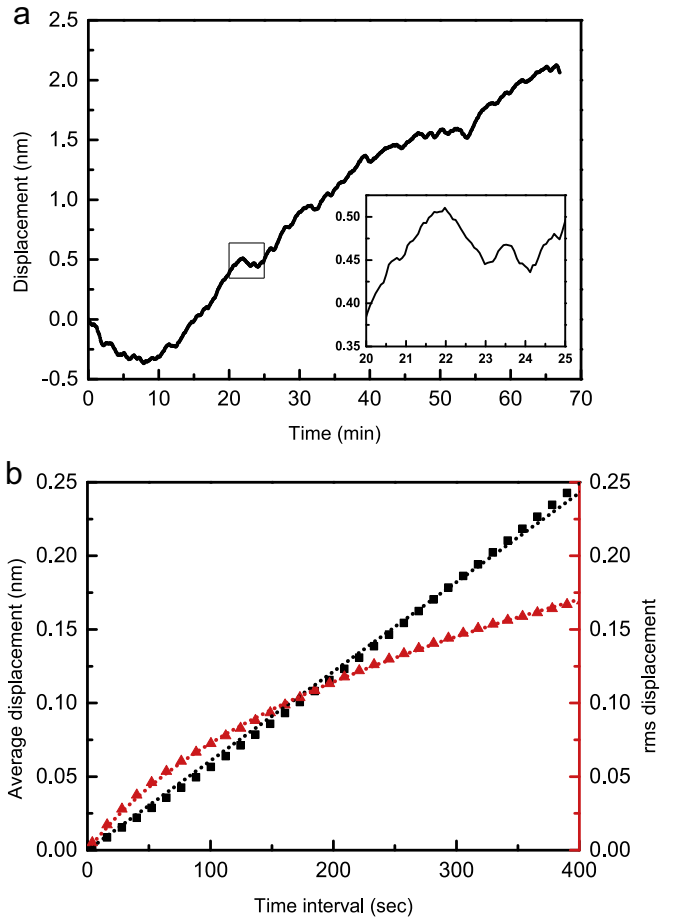

Fig. 4. (a) Fringe displacement as a function of time, measured using 1000 blank holograms over a timespan of 67 min. The inset reveals the timescale of the random motion. (b) The average (black squares) and rms (red triangles) displacement for time intervals up to 400 s. The theoretical models are overlaid (dotted lines). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

a larger effective signal by improving the visibility (better spatial coherence). Hence there exists an optimum signal strength that produces the largest effective signal and the smallest phase error. For signal strengths less than the optimum the phase measurement becomes *noise limited*, whereas for signals greater than the optimum the measurement becomes *coherence limited*.

5.2. Time-dependent visibility

The above results concerned the spatial coherence of the wavefield via the signal strength. In particular, the exposure time was kept constant. However, the visibility has a time-dependent component that is governed by the instability of the instrument. Here we have investigated the fringe displacement as a function of time, via the acquisition of 1000 blank holograms over 67 min. The fringe displacement was tracked via the phase of the sideband's Fourier component. The result is shown in Fig. 4. Firstly, we observe that the fringes moved initially in a negative direction and then a positive direction. Secondly, the movement consists of smaller random fluctuations residing on a more slowly varying, near-linear movement. The time scale of the random component is seen to be on the order of one minute (inset of Fig. 4(a)). In passing, we note that similar stochastic behaviour has recently been observed for the aberrations in high-resolution TEM [22].

Statistical analysis of the fringe displacement is shown in Fig. 4 (b), which plots the average and the root-mean-square (rms) displacement as a function of the time interval. The average displacement is well modelled by a linear trend with a velocity of 0.6 pm/s. The rms displacement is well modelled by the Langevin theory (Eq. (9)) with an average time and distance between "collisions" of $t_c \simeq 36$ s and $d_c \simeq 0.038$ nm, respectively.

The fitted values of t_c and d_c allow us to estimate the optimum exposure time based on Fig. 2. For example, for a fringe spacing of 0.08 nm (applicable to our primary data), the optimum exposure time is $(d/d_c)t_c = 18$ s. For long exposure times (greater than about 100 sec), the linear component becomes important (or dominant) and therefore has to be taken into account.

5.3. Normalised effective signal and phase error

The previous subsections have established the general trends of spatial coherence and instrument instabilities. In this section, we present an analysis of our primary dataset, where the signal level and the exposure time are varied systematically. Here we also introduce the concept of normalised effective signal which has the benefit of revealing a universal behaviour with respect to the exposure time. It remains to fit our spatial coherence model, which is done below.

Fig. 5 shows the phase errors obtained from our primary dataset. It is observed that the minimum phase error obtained decreases with increasing exposure time up to approximately 16 s.

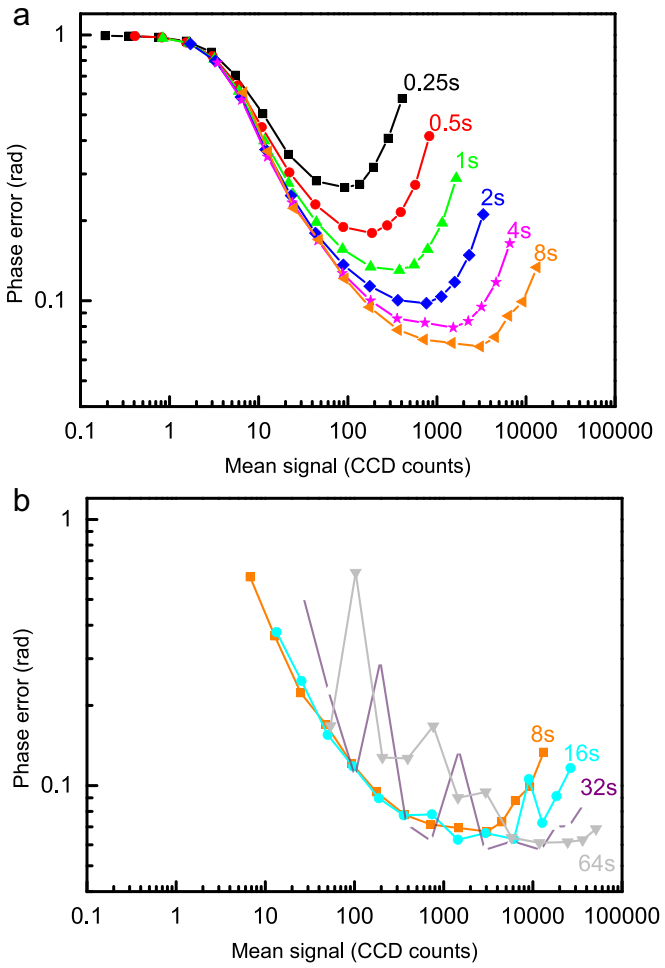

Fig. 5. Phase error as a function of mean signal for exposure times of 0.25–8 s (a) and 8–64 s (b). For each exposure time, the value of the mean signal was controlled by the C2 lens excitation.

For greater exposure times, the minimum phase error does not exhibit a further decrease, and instead the behaviour of the phase error becomes erratic. This observation is consistent with the predicted optimum exposure time of 18 s stated in the previous section: for exposure times of 32 and 64 s, which are well in excess of the optimum, the fringe displacement becomes, on average, excessively large, ultimately leading to greater phase errors (smaller values of the phase error are occasionally observed, consistent with the stochastic nature of the fringe displacement in Fig. 4(a)).

The final step in our analysis is to fit the model for the spatial coherence $S(\hat{N})$ given by Eq. (5). The fitting of $S(\hat{N})$ was performed on the visibility V measured from the primary dataset with a 2 s exposure time, using a least-squares criterion. This exposure time was chosen to obtain good statistics while minimising the effects of instrument instabilities. The fitted parameters were $\alpha = 2.0 \times 10^{-5}$ s$^{1/2}$, and $\beta = 2.5 \times 10^{-10}$ s. This fitting procedure also refines a value for the time- and dose-independent part of the visibility $V_0 = 0.27$.

Having fitted all of the parameters in our model, we now compare it with the results obtained from our primary dataset. To facilitate the comparison, we consider the normalised effective signal ($\hat{N}_{\mathrm{eff}} = N_{\mathrm{eff}}/t$) as a function of the normalised signal ($\hat{N} = N/t$), which reveals the universal trend for all exposure times. Fig. 6 shows the comparison for selected exposure times. We see that for exposure times of 0.5 and 2 s, $\hat{N}_{\mathrm{eff}}$ remains virtually unchanged, reflecting that the time-dependent visibility $V(t) \approx 1$ for these times (see Fig. 2). Further increasing the exposure time eventually results in an overall, but somewhat random, decrease of $\hat{N}_{\mathrm{eff}}$, reflecting the decrease in $V(t)$. The measured values of $\hat{N}_{\mathrm{eff}}$ are well-reproduced by the fitted model. For 32 s, which is well beyond the 18 s optimum for this condition, some discrepancy between the measured values and the model is observed, but this is expected given the highly stochastic nature of the instabilities at such long times.

Also shown in Fig. 6 are the corresponding phase errors ($\Delta\phi$). Unlike $\hat{N}_{\mathrm{eff}}$, the phase errors immediately improve (decrease) with exposure time because they are governed by the effective signal N_{eff} (not the normalised one). At and beyond the 18 s optimum, the phase error shows no further improvement (compare 8 and 32 s in Fig. 6), but instead degrades in a random fashion, in line with the stochastic behaviour of $\hat{N}_{\mathrm{eff}}$ at these times.

Crucial to the present work, the maximum values of $\hat{N}_{\mathrm{eff}}$ indeed correspond to minimum values of $\Delta\phi$ (as indicated by the dashed vertical lines in Fig. 6). This confirms that an analysis of $\hat{N}_{\mathrm{eff}}$ is sufficient for predicting the optimum $\hat{N}$ and t, which greatly simplifies the analysis. (On the other hand, it is not possible from a knowledge of $\hat{N}_{\mathrm{eff}}$ alone to predict the actual value of the minimum phase error $\Delta\phi$, since that value also depends on the detector gain G, as well as other secondary factors [16].)

5.4. Prediction and verification of optimum conditions

The model presented above can be used to identify the optimum dose rate and exposure time for a given fringe spacing d, interference width w and pixel size δ. Once the optimum dose rate $\hat{N}$ and exposure time t are determined for fixed parameters d_0, w_0 and δ_0, we can use them in conjunction with Eq. (14) and Fig. 2b to predict the optimum dose rate and exposure time for other sets of parameters. Table 1 exemplifies this methodology.

Qualitatively speaking, a larger fringe spacing (obtained by decreasing the biprism voltage) permits a more intense electron beam (since the spatial coherence requirements are less stringent) and a longer exposure time (since the stability requirements are less stringent). A higher magnification (with all other parameters fixed) implies a reduced optimum dose simply because of the reduced flux of electrons at the detector. To verify these predictions, Table 1 also lists the independently measured optimum dose for each condition. It can be seen that good agreement is obtained between the theoretical predictions and experimental

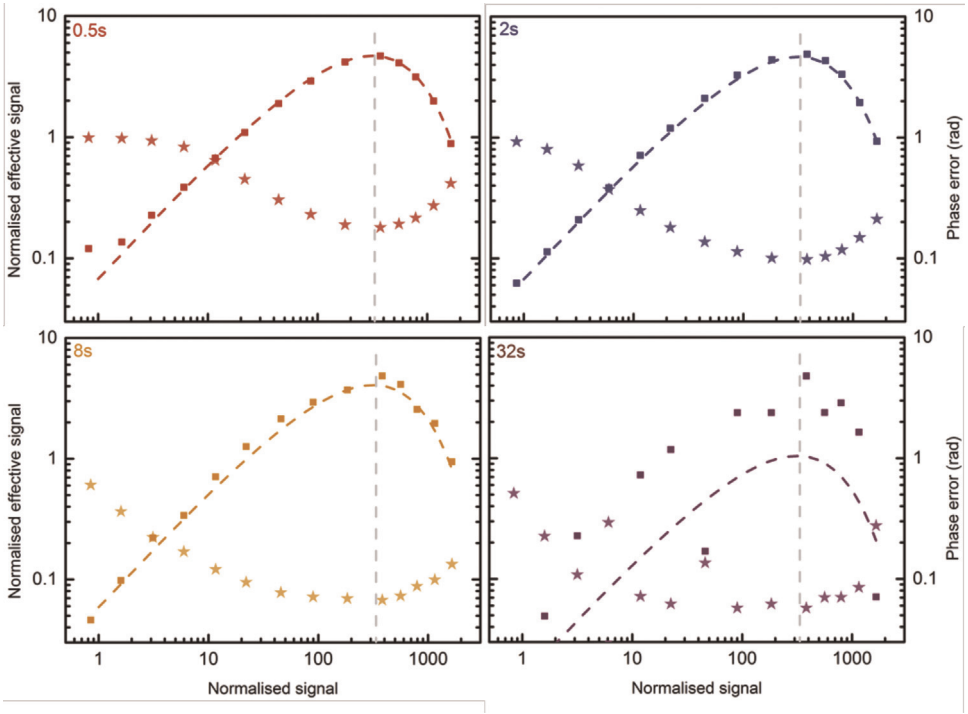

Fig. 6. The normalised effective signal $\hat{N}_{\mathrm{eff}}$ (squares) and the corresponding phase error $\Delta\phi$ (stars) as functions of the normalised signal $\hat{N}$, plotted for selected exposure times (indicated top-left of each graph). Also shown are predictions for $\hat{N}_{\mathrm{eff}}$ (dashed lines) based on fitting described in the text. The grey vertical dashed lines indicate the optimum normalised signal.

Table 1
Predicted optimum doses N_{opt} and exposure times t_{opt} for a selection of experimental conditions. The last column lists independently measured values of N_{opt} (except that marked "*", which is not an independent measurement).

Mode	M	δ (nm)	V_{BP} (V)	d (nm)	w (nm)	N_{opt}	t_{opt} (s)	N_{opt} Meas.
Standard	450k	0.0192 ± 0.001	150	0.084 ± 0.002	31 ± 0.2	316	18	315 ± 64*
Standard	940k	0.0088 ± 0.002	70	0.175 ± 0.002	16.24 ± 0.2	242	46	225 ± 45
Standard	1950k	0.0042 ± 0.003	65	0.176 ± 0.006	16.37 ± 0.2	54	46	60 ± 15
Lorentz	62k	0.225 ± 0.01	150	0.98 ± 0.025	364 ± 9	316	18	368 ± 80
Lorentz	62k	0.225 ± 0.01	87	1.666 ± 0.06	176 ± 8	1349	45	1200 ± 200

measurements. The error in the measured optimum dose was chosen to correspond to a 5% reduction in the effective signal. This accounts for the fact that the intensity varies slowly around the optimum, so that electron doses falling within the error satisfies the optimum dose condition to a good level of accuracy.

The above methodology also applies to Lorentz mode. This mode produces a larger field of view because of the reduced magnification between the specimen and the biprism. Notwithstanding this, given the results in Fig. 3, we anticipate that Lorentz mode does not significantly affect the spatial coherence envelope $S(\hat{N})$. This prediction is verified experimentally in Fig. 7 where the excitations of the post-biprism lenses were purposely kept identical. Furthermore, we do not anticipate any significant change in the instrument instabilities responsible for the time-dependent visibility $V(t)$. Hence our analysis in standard mode can be directly applied to Lorentz mode, provided that the change in specimen magnification is accounted for. The last two rows in Table 1 apply to Lorentz mode.

6. Extension to elliptical illumination

While the methodology described above assumed round illumination, it is easily extended to the case of elliptical illumination. The only aspect of the model requiring alteration is the spatial coherence envelope, which becomes

$$S(\hat{N}) = e^{-\alpha'(bw/\delta^2)\hat{N} - \beta'(bw/\delta^2)^2\hat{N}^2}. \tag{15}$$

where b is the semi-minor axis of the ellipse (i.e., $2b$ is the size of the illumination measured parallel to the biprism wire). Note the appearance of $\hat{N}$ and $\hat{N}^2$ in Eq. (15) (as opposed to $\hat{N}^{1/2}$ and $\hat{N}$ in Eq. (5)), which implies that the parameters α' and β' differ from α and β (hence the primes). Despite these differences, the underlying effective source model remains the same.

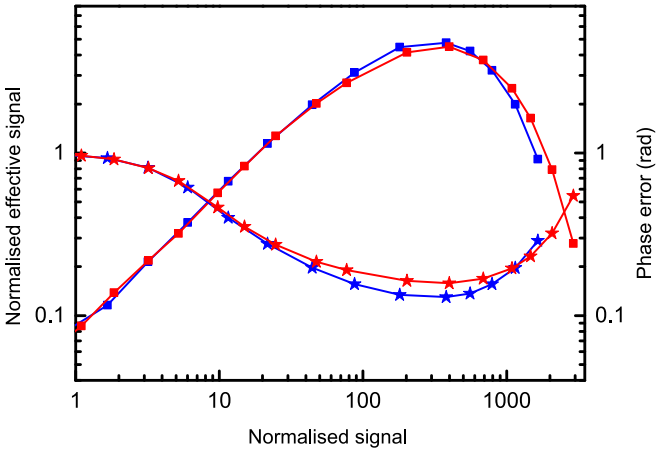

Fig. 7. Comparison of standard mode and Lorentz mode for (virtually) identical excitations of the post-biprism lenses. The graph shows the normalised effective signal (squares) and the phase error (stars) for a standard mode magnification of 450k (blue) and a Lorentz mode magnification of 67k (red), plotted as a function of normalised signal. The biprism voltage was 150 V. The exposure time was 1 s. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

With Eq. (15), the optimisation proceeds almost exactly as before, except that the set of fixed parameters d, w and δ now includes b. In practice, b is chosen to be as small as possible while still providing uniform illumination across the field of view. Instead of Eq. (12), the optimum dose rate for elliptical illumination is given by

$$\max_{\hat{N}} N_{\text{eff}}(\hat{N}, t) \Longrightarrow \hat{N} = \frac{\delta^2}{bw}\frac{\sqrt{\alpha^2 + 4\beta} - \alpha}{4\beta}. \tag{16}$$

Instead of Eq. (14), the optimum dose rate is scaled for other conditions according to

$$\hat{N} = \frac{b_0 w_0/\delta_0^2}{bw/\delta^2}\hat{N}_0. \tag{17}$$

Experimentally, since b is fixed, the variation and optimisation of $\hat{N}$ is now controlled by "stretching" the illumination across the biprism, using, for example, a combination of the C2 lens and the condenser stigmators.

7. Conclusions

To facilitate the optimisation of phase errors in electron holography, we have established a relatively simple and yet sufficiently accurate model that can be fitted using a minimal experimental dataset. Once fitted, the model is capable of predicting the optimum dose rate and exposure time for any given combination of biprism voltage and magnification for a single exposure, either in standard mode or Lorentz mode. The optimum dose rate and exposure time produce the smallest-possible phase error for the given conditions on a particular instrument, not explicitly taking into account specimen considerations. In the presence of specimen drift and/or dose-dependent specimen damage, the optimum dose rate remains entirely valid, although the exposure time may need to be reduced accordingly.

Experimental studies were undertaken to verify the models of the spatial coherence and instrument instabilities that are required for the optimisation. We found that the commonly used Gaussian model is not suitable to describe the spatial coherence, and instead a bivariate Cauchy distribution convoluted with a Gaussian distribution is better suited. The fringe movement due to instabilities is well-modelled by the Langevin theory of Brownian motion, which improves upon previous models since it is applicable to the practical range of exposure times used in experiments.

We conclude with a recipe for obtaining optimum dose and exposure time using our model: in addition to a knowledge of the biprism width, two sets of blank holograms are required, each set using the same biprism voltage and magnification (i.e., the primary values of d, w and δ). The first set is acquired as a function of

intensity (e.g., by changing the C2 lens) using a short exposure time to minimise instabilities (e.g., 1 s). The fringe visibilities in this set are used to fit the spatial coherence envelope $S(\hat{N})$. The second set consists of a time series, acquired over a long timespan (e.g., 1 h), again using a short exposure time for each hologram. The fringe positions in this set are used to fit the time-dependent visibility $V(t)$. Once the model has been fitted, it can be used to predict the optimum dose rate and exposure time for other conditions. Alternatively, the optimum dose rate for the initial condition can be scaled to other biprism voltages and magnifications using Eq. (14).

Acknowledgements

S.L.Y.C. acknowledges fruitful discussions with Dr. Juri Barthel and Dr. Emrah Yücelen. S.L.Y.C., C.D. and C.B.B. acknowledge helpful discussions with Dr. Thomas Duden and Prof. Guilio Pozzi. The authors gratefully acknowledge the European Commission for an Advanced Grant.

References

[1] M. Linck, B. Freitag, S. Kujawa, M. Lehmann, T. Niermann, Ultramicroscopy 116 (2012) 13.
[2] B. Freitag, G. Knippels, S. Kujawa, P. C. Tiemeijer, M. Van der Stam, D. Hubert, C. Kiesielowski, P. Denes, A. Minor, U. Dahmen, in: Proceedings of EMC: Instruments and Methods, vol. 1, 2008, p. 55.
[3] D.C. Joy, Y.S. Zhang, T. Hashimoto, R.D. Bunn, Ultramicroscopy 51 (1994) 1.
[4] M. Lehmann, Ultramicroscopy 100 (2004) 9.
[5] D. Cooper, R. Truche, P. Rivallin, J.M. Hartmann, F. Laugier, F. Bertin, A. Chabli, J. L. Rouviere, Appl. Phys. Lett. 91 (2007) 143501.
[6] P. Grob, D. Bean, D. Typke, X. Li, E. Nogales, R.M. Glaeser, Ultramicroscopy 133 (2013) 1.
[7] E. Voelkl, Ultramicroscopy 110 (2010) 199.
[8] R.A. McLeod, M. Bergen, M. Malac, Ultramicroscopy 141 (2014) 38.
[9] T. Niermann, M. Lehmann, Micron 63 (2014) 28.
[10] Q. Ru, J. Endo, T. Tanji, A. Tonomura, Appl. Phys. Lett. 59 (1991) 2372.
[11] K. Yamamoto, I. Kawajiri, T. Tanji, M. Hibino, T. Hirajama, J. Electron Microsc. 49 (2000) 31.
[12] T. Suzuki, S. Aizawa, T. Tanigaki, K. Ota, T. Matsuda, A. Tonomura, Ultramicroscopy 118 (2012) 21.
[13] H. Lichte, K.H. Herrmann, F. Lenz, Optik 77 (1987) 135.
[14] W.J. de Ruijter, J.K. Weiss, Ultramicroscopy 50 (1993) 269.
[15] A. Harscher, H. Lichte, Ultramicroscopy 64 (1996) 57.
[16] F. Röder, A. Lubk, D. Wolf, T. Niermann, Ultramicroscopy 144 (2014) 32.
[17] H. Lichte, Ultramicroscopy 108 (2008) 256.
[18] M. Born, E. Wolf, Principles of Optics, 7th Ed., Cambridge University Press, Cambridge, UK, 1999.
[19] C. Dwyer, R. Erni, J. Etheridge, Ultramicroscopy 110 (2010) 952.
[20] C. Maunders, C. Dwyer, P. Tiemeijer, J. Etheridge, Ultramicroscopy 111 (2011) 1437.
[21] J. Verbeeck, A. Beche, W. Van den Broek, Ultramicroscopy 120 (2012) 35.
[22] J. Barthel, A. Thust, Ultramicroscopy 134 (2013) 6.