Ultramicroscopy 50 (1993) 269–283
North-Holland

ultramicroscopy

Detection limits in quantitative off-axis electron holography

W.J. de Ruijter [1] and J.K. Weiss
Center for Solid State Science, Arizona State University, Tempe, AZ 85287, USA

Received 11 February 1993; in final form 1 June 1993

The phase of an electron wave is altered by electric and magnetic fields as it passes through a specimen. This phase change can be accurately quantified from off-axis electron holograms acquired using a slow-scan CCD camera, and small changes can be observed over small dimensions. Expressions for the precision of the phase estimate, which is limited by shot noise, have been developed. These include most of the real experimental parameters. It is found that the typical precision of practical phase measurements is better than $\pi/100$ for spatial resolutions of 1–3 nm, in good agreement with the theoretical optimal phase precision. In order to attain such small errors the effects of geometric distortion, which can introduce phase differences of up to π, must be carefully corrected.

1. Introduction

In conventional transmission electron microscopy only the amplitude of the image wave is recorded and the phase is lost. Electron holography enables measurement of the phase thus accommodating a new class of experiments aimed at direct determination of properties of electric and magnetic fields in specimens [1–7]. Furthermore, remarkable progress has been made in recent years in the application of electron holography for exit-surface wave reconstruction by the removal of objective lens aberrations from the complex image wave [8].

Accurate quantitative analysis of information contained in electron holograms has been difficult in the past due to limitations caused by the recording medium and the procedures used to extract the phase. Until recently, phase retrieval was based on a reconstruction procedure implemented on the optical bench. This technique does not allow for easy quantitative analysis and, furthermore, measurement of very small phase differences, as encountered, for example, in studies of monatomic surface steps, is complicated. For example, it has been shown that detection of small differences (of the order of $2\pi/50$) is possible with a phase difference amplification method which involves a repeated application of optical reconstruction that, while effective, is extremely tedious [2,9]. In recent years holographic reconstruction has been implemented on the computer, which necessitated digitization of the photographic plate [8]. It has already been demonstrated that digital reconstruction makes visualization of phase shifts on the order of $2\pi/100$ possible [10]. Unfortunately, quantitative analysis was still not feasible, in part due to problems associated with the photographic recording medium. The effect of the non-linear response of photographic material strongly affects holographic reconstruction, and therefore complicated linearization techniques must be applied [11]. Furthermore, the projector lens system causes geometric distortion which distorts absolute phase measurements: this distortion is very difficult to correct accurately from holograms recorded photographically.

[1] Correspondence to: Dr. W.J. de Ruijter. Present address: Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA 94143-0448, USA.

Our aim in this paper is to show that the process of holographic reconstruction can be considerably simplified, and the detection limits and measurement accuracy can be dramatically improved by application of the slow-scan CCD camera. This device records digitized images with a large number of pixels (1024 × 1024), negligible geometric distortion, perfect positional stability, excellent linearity and approaches single-electron detection [12,13]. Digital acquisition of electron holograms with the slow-scan CCD camera removes the necessity for tedious linearization procedures, facilitates simple geometric distortion correction, and creates the possibility of on-line digital reconstruction. Quantitative analysis and visualization of the reconstructed phase can now be performed with established digital image-processing techniques, and elementary procedures such as averaging and gray-scale enhancement give dramatic improvements compared with optical reconstruction techniques.

We present a theoretical analysis based on statistical parameter estimation which establishes the ultimate precision of phase measurements as a function of important experimental parameters such as electron dose and biprism voltage. Furthermore, attention has been given to experimental design considerations in order to optimize this precision. The measurement techniques presented in this paper have been used in the companion papers on applications of electron holography to the study of interfaces [14] and to measure absolute mean inner potential [15].

2. Theory

2.1. Electron holography

Interference between the object wave and a reference wave is made possible by means of a biprism located between the back focal plane of the objective lens and the object plane of the intermediate lens. The geometry is illustrated in fig. 1. The amplitude of the image wave is present in the hologram as a modulation of the intensity of the holography fringes, whereas the phase is visible through local fringe shifts. It is useful to

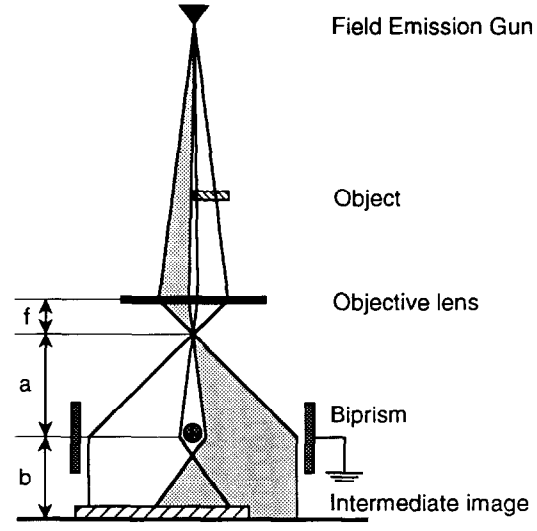

Fig. 1. Principle of electron holography. The reference wave (shaded) and the object wave interfere due to electrostatic biprism.

consider the Fourier transform of the hologram intensity which is given by

$$F[I_{\mathrm{hol}}(\boldsymbol{r})] = \delta(\mathbf{0}) + F(\psi^*\psi) + F(\psi) \otimes \delta(\boldsymbol{g}_{\mathrm{h}}) + F(\psi^*) \otimes \delta(-\boldsymbol{g}_{\mathrm{h}}), \quad (1)$$

where $\boldsymbol{r}$ denotes position in the image plane, $\boldsymbol{g}_{\mathrm{h}}$ is the spatial frequency of the holography fringes, δ is the Dirac delta function, $\otimes$ denotes convolution and F denotes Fourier transformation, ψ is the image wave given by

$$\psi(\boldsymbol{r}) = A(\boldsymbol{r}) \exp[\mathrm{i}\phi(\boldsymbol{r})], \quad (2)$$

with $A(\boldsymbol{r})$ the amplitude and $\phi(\boldsymbol{r})$ the phase. The first terms on the right-hand side of eq. (1) represent the information also found in the diffractogram of a bright-field image, namely the Fourier transform of the squared image intensity. Additionally, the Fourier transform of the hologram contains information in two sidebands, centered around $-\boldsymbol{g}_{\mathrm{h}}$ and $\boldsymbol{g}_{\mathrm{h}}$, which allow direct retrieval of the complex image wave.

2.2. Precision of phase measurement

Due to coherence requirements, holograms are typically recorded at electron doses of 100–200

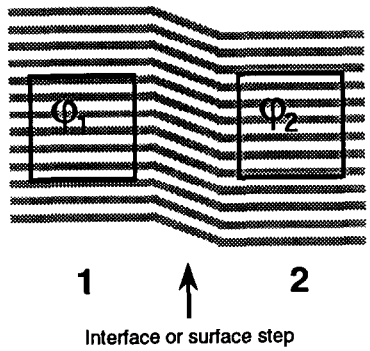

Fig. 2. Bending of holography fringes at interface or surface step. Reconstruction of phase from electron hologram is equivalent to estimation of phase of holography fringes, for example, ϕ_1 and ϕ_2, in regions 1 and 2.

el/px. Since holograms contain noise which limits the measurement precision, the precision of the phase measurement must be evaluated using statistical analysis. We need to measure the phase of two-dimensional holography fringes (which are two-dimensional sinusoids), as illustrated in fig. 2, that are disturbed by noise.

Noise in an electron hologram is due to single-electron events. The shot noise is white and Poisson-distributed with a standard deviation of $\sigma = \sqrt{N_e}$ where N_e is the average number of detected electrons per pixel. The amplitude of the fringes can also be expressed in terms of electron dose as $A = VN_e$ where V is the fringe visibility; a definition commonly used in light-optics [16]. The achievable precision for estimation of the phase is established in appendix A as:

$$\mathrm{var}\left[\hat{\phi}\right] \geq \frac{14}{V^2 N_t}, \tag{3}$$

where $\hat{\phi}$ is an estimator for the quantity ϕ and N_t denotes the total electron dose in the measured area. Notice that the precision is not related to the spacing of the holography fringes. The bound (3) can alternatively written as $\mathrm{var}[\hat{\phi}] \geq 7/\mathrm{SNR}$ with signal-to-noise ratio $\mathrm{SNR} = V^2 N_t/2$.

The minimum variance bound (3) is larger than the bound given in refs. [17,18]: $\mathrm{var}[\hat{\phi}] \geq 2/V^2 N_t$. This bound was calculated for one-dimensional signals with the assumption that the sinusoidial fringe spacing is a-priori known, and the fringe amplitude is unknown (in this case the extension to two dimensions is trivial). However, in electron holography the x and the y components of the spatial frequency vector are never a-priori known in practice, and although they are of no specific interest, they significantly influence the precision of the phase estimate since these quantities are unknown. The calculations in appendix A reconfirm the bound for the phase estimate given in refs. [17,18], and show that inclusion of the spatial frequency in the x and the y direction as unknowns increases the minimum variance bound 7 times.

The practically achievable variance bound will be larger than (3) due to loss of signal-to-noise ratio associated with detection of the electron hologram. Fortunately, the slow-scan CCD camera is a near-ideal electron detector, and the loss of precision is small. The slow-scan CCD camera can be best characterized using the transfer efficiency (TE), which is given by [19]:

$$\mathrm{TE}(\boldsymbol{g}) = \mathrm{SNR}_{\mathrm{out}}(\boldsymbol{g})/\mathrm{SNR}_{\mathrm{in}}(\boldsymbol{g}), \tag{4}$$

with $\mathrm{SNR}_{\mathrm{out}}(\boldsymbol{g})$ and $\mathrm{SNR}_{\mathrm{in}}(\boldsymbol{g})$ the signal-to-noise ratios of the final digital image in the computer and the original electron hologram, respectively. Signal-to-noise ratio is defined as the ratio of the power of the signal and the power of the noise at spatial frequency $\boldsymbol{g}$. For the ideal camera the transfer efficiency is unity over all spatial frequencies. However, this value is not achieved in practice, due to backscattering of a percentage of primary electrons incident on the scintillator [19]. For the electron doses typical in electron holography (100–200 el/px) the transfer efficiency proves to be a constant which is $\mathrm{TE} \approx 0.8$ for the Gatan 679 CCD camera equipped with thin YAG scintillator [19]. With eq. (4) the minimum variance bound (3) can be rewritten as:

$$\mathrm{var}\left[\hat{\phi}\right] \geq \frac{1}{\mathrm{TE}} \frac{14}{V^2 N_t}. \tag{5}$$

The fact that the transfer efficiency of the camera is independent of the spatial frequency of the holography fringes might at first seem surprising,

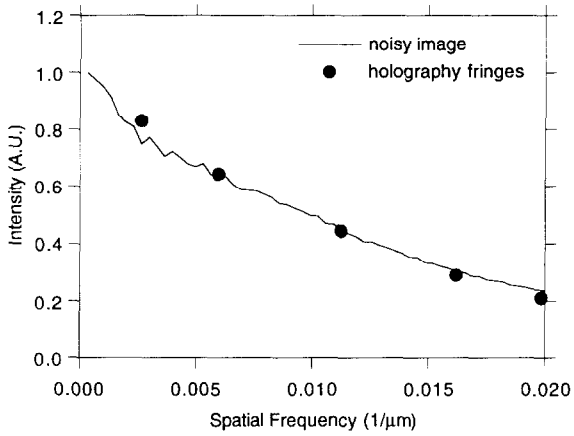

Fig. 3. Amplitude of holography fringes at five different frequencies compared with the square root of the power spectrum of the shot noise, as measured from a division of two images recorded with uniform illumination without specimen. The amplitude of the holography fringe with lowest spatial frequency is used to scale the measured amplitudes to the power spectrum of the shot noise.

because fringes with higher spatial frequency are attenuated considerably due to the modulation transfer function of the CCD camera, which, for our Gatan 679, is about 0.2 at the Nyquist frequency [13]. Consider, however, that the shot noise in the hologram should be attenuated by exactly the same amount. Therefore, if the electron dose is not too small, such that the shot noise due to primary electrons incident on the scintillator dominate the noise contribution originating from electronic read-out noise and the spread in CCD well-electron generation, the transfer efficiency should be independent of the spatial frequency [19]. This is illustrated by the measurement results presented in fig. 3, showing a comparison of the square root of the power spectrum of the shot noise and the amplitude of holography fringes imaged at different microscope magnifications.

2.3. Experimental design considerations

The precision (5) is determined by the product V^2N_t which is dependent upon several variables. Tuning those variables to optimize the measurement precision is known as experimental design [20]. This section introduces important physical parameters, determines their influence on the precision of the phase measurement, in turn leading to recommendations for optimal experimental conditions.

The variables which determine the total number of electrons N_t incident in an area A_R follow from [21]

$$N_t = \frac{\pi^2 \beta R_s^2 \alpha_0^2 \tau}{e} \frac{4A_R}{\pi d_1 d_2}, \tag{6}$$

where β is the gun brightness, R_s denotes the Gaussian source size projected at the specimen plane, α_0 is the convergence half angle of the illumination, τ is the measurement time, and d_1, d_2 are the major and minor axes, respectively, of the astigmatic illumination. The fringe visibility V is highly dependent on the length d_1 of the illuminated region in the direction perpendicular to the holography fringes, and can be described by [22,23]

$$V = \exp\left[-\left(\frac{2\pi R_s \alpha_0 D}{\lambda d_1}\right)^2\right], \tag{7}$$

where D is the distance between areas at the specimen plane which interfere behind the biprism.

The experimental parameters which are easiest to adjust are the illumination diameters d_1 and d_2. The product V^2N_t can be maximized with respect to d_1 using eqs. (6) and (7), as is illustrated in fig. 4, with the result

$$d_1 = \frac{4\pi R_s \alpha_0 D}{\lambda}. \tag{8}$$

The width d_2 must be large enough to illuminate width D evenly because for large convergence angles ($\alpha_0 > 10$ mrad) spherical aberration of the upper half of the objective lens, $C_s^{(u)}$, may dominate the intensity distribution on the sample [24]. Since the spherical aberration caustic has a radius of about $r_s = C_s^{(u)}\alpha_0^3$, as long as $d_2 \gg r_s$, which will generally be true for $d_2 = kD$ with

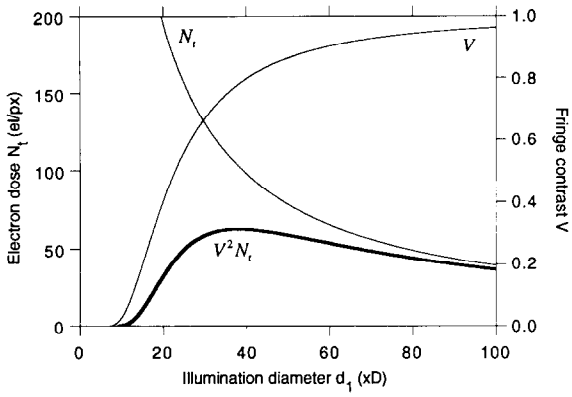

Fig. 4. Optimization of V^2N_t with respect to illumination diameter d_1. With experimental conditions given in tables 1 and 2 the optimal choice for d_1 is $38D$, D being the distance of interference distance scaled to the specimen, resulting in an electron dose of 200 el/px and fringe visibility of 78%.

$k > 1$, the specimen will be evenly illuminated. With these values for d_1 and d_2, it follows that:

$$N_t = \frac{\beta R_s \alpha_0 \lambda}{e} \frac{A_R \tau}{kD^2}, \tag{9}$$

$$V = \exp[-1/4]. \tag{10}$$

The width D can be expressed in terms of the biprism voltage U_B as:

$$D = \frac{2b\gamma_0}{M_{obj}} U_B, \tag{11}$$

where b is the distance between the biprism and the intermediate image plane, M_{obj} is the objective lens magnification, and γ_0 is a constant depending on the accelerating voltage of the microscope [8]. Following eqs. (5), (9) and (11) the variance for the phase estimate is proportional to U_B^2/τ, which implies that the precision (5) can be optimized by choosing large τ and small U_B. Unfortunately, these parameters cannot be chosen arbitrarily.

The integration time τ is limited by the stability of the microscope, specifically the specimen stage and the biprism mount. Drift of the specimen stage leads to loss of spatial resolution, and since standard specimen holders are often not rated at better than 1 nm/min drift, the acquisition time must usually be kept between 1 and 5 s. Similarly, if the biprism position drifts during acquisition, the fringe contrast will be reduced, but since the objective lens magnification is high and the biprism lies in the image plane of the objective, there is considerably more tolerance for biprism drift.

The choice of the biprism voltage U_B is limited because the width of the field of view (11) and the interference fringe spacing s

$$s = \frac{\lambda f}{2\gamma_0 a U_B}, \tag{12}$$

where f is the focal length of the objective lens and a is the distance between the back focal plane of the objective lens and the biprism, are both dependent upon U_B. In the following we will describe requirements for D and s, which lead to a specific choice for U_B.

The width of the holographic interference band (11) at the detector plane should be about the same size as the detector. The width of this interference region is slightly less than D due to vignetting of the biprism, resulting in:

$$MD = wN\Delta_c, \tag{13}$$

where M is the microscope magnification, N is the number of pixels across the CCD array, Δ_c is the camera pixel size, and w is a constant of value generally between 1 and 2. In addition, it is necessary to make the fringe spacing s (12) greater than the Nyquist limit 2Δ, where $\Delta = \Delta_c/M$. In order to prevent overlap of the sidebands of the hologram with the central diffractogram, it is necessary to locate the sidebands an appreciable distance away from the center, leading to a choice

$$Ms = n\Delta_c, \tag{14}$$

where n is a number in general between 3 and 5. Specific choices for w and n lead to unique magnification and biprism voltage settings, which are found by substitution of eqs. (11) and (12) in eqs. (13) and (14), respectively:

$$M = \Delta_c \left[\frac{M_{obj} a}{\lambda f b} wNn \right]^{1/2}, \tag{15}$$

$$U_B = \frac{1}{2\gamma_0} \left[\frac{M_{obj} \lambda f}{ab} \frac{wN}{n} \right]^{1/2}. \tag{16}$$

Table 1
Typical values for experimental parameters for electron holography on Philips EM400ST-FEG with Gatan 679 CCD camera

Parameter	Value	Parameter	Value
λ	0.0037 nm	R_s	0.8 nm
γ_0	1.96×10^{-6} rad/V	α_0	14 mrad
e	1.602×10^{-19} C	N	1024
M_{obj}	55	Δ_c	24 μm
f	1.5 mm	TE	0.8
a	61 mm	τ	1 s
b	16 mm	n	4
β	4×10^{12} A/m^2·sr	k	2
		w	1.5

An expression for the for the optimal variance of the phase estimate (5) follows from eqs. (9) and (10) combined with eqs. (11) and (16):

$$\mathrm{var}\left[\hat{\phi}\right] = \frac{14\exp[1/2]}{\mathrm{TE}}\,\frac{e}{\beta R_s \alpha_0 \tau}\,\frac{fb}{M_{obj} a}\,\frac{wkN}{nA_R}. \tag{17}$$

Typical values of experimental parameters are shown in table 1. The design rules outlined above, combined with the experimental parameters listed in table 1, result in the optimal experimental conditions listed in table 2. The optimal standard deviation in the measured phase follows from eq. (17) by substituting values listed in table 1 and is of the order of $\pi/100$ radians for small measurement areas of the order of $A_R = 1$ nm^2.

3. Experimental holography

3.1. Digital acquisition of holograms

The off-axis electron holograms were produced in a Philips EM400ST-FEG electron microscope equipped with a thermally assisted field emission gun [25]. The holograms were recorded on a Gatan 679 slow-scan CCD camera equipped with a 1024×1024 pixel detector. The biprism was a thin glass fiber (~ 0.5 μm diameter) coated with a thin layer of gold which was mounted in one of the selected area aperture positions. The

Table 2
Optimal experimental conditions for electron holography on Philips EM400ST-FEG with Gatan 679 CCD camera

Parameter	Value	Parameter	Value
d_1	$38D$	M	366000
d_2	$2D$	Δ	0.066 nm
D	100 nm	U_B	88.4 V
s	0.26 nm	V	78%
		N_t	50000 el/nm^2

biprism was electrically isolated from the microscope column, and was connected to an external high-voltage battery source which could produce variable potentials of up to 180 V on the biprism. In order to facilitate electron holography, the electronics of the microscope had to be slightly altered to accommodate a higher current in the first intermediate lens in order that the first intermediate image was formed below the biprism.

The CCD images were recorded using a Macintosh IIfx computer equipped with 32 MB of memory, a two-page monitor and a 20 MFlop Mercury MC3200 floating point array processor. The software delivered with the slow-scan CCD camera (DigitalMicrograph) allowed user-oriented expansion through custom functions. We

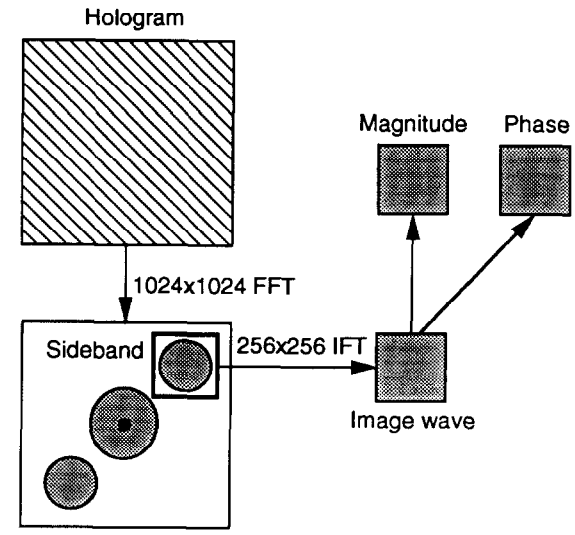

Fig. 5. Reconstruction of magnitude and phase of the complex image wave. FFT is fast forward Fourier transform and IFT is fast inverse Fourier transform.

developed as set of custom functions which enabled us to perform fast on-line and off-line holographic reconstructions.

Due to coherence requirements, a dose rate of approximately 50000 el/nm^2 · s was incident on the specimen. In order to reduce the effects of specimen and/or biprism drift, exposure times of about 1 s were typically used. With magnifications of about 300k ×, at which 0.3 nm interference fringes corresponded to a distance of $4\Delta_c$ at the CCD detector, this dose rate corresponded to about 150 el/px on the detector.

The fiber-optic coupling which transferred photons from the YAG scintillator to the CCD

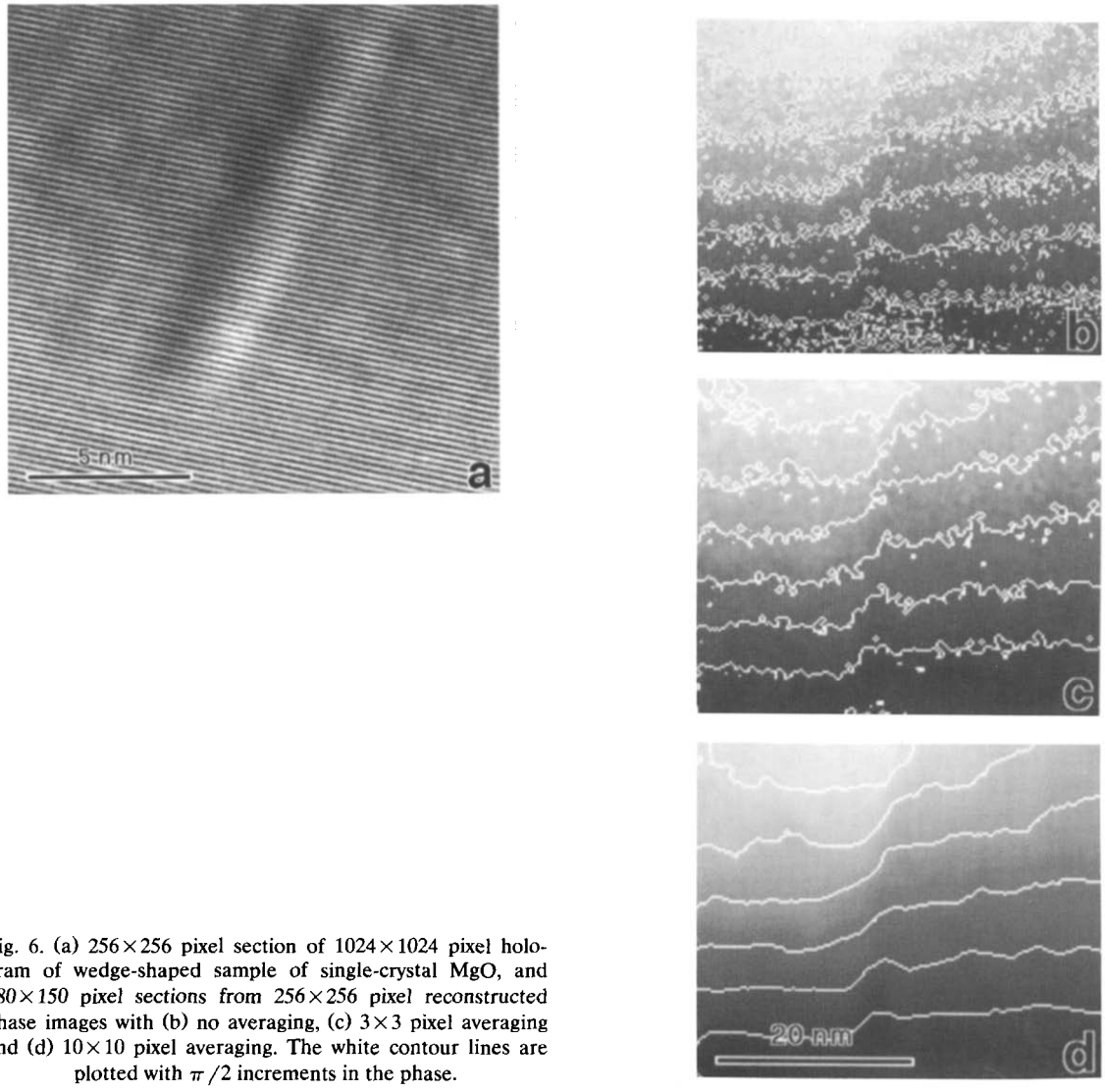

Fig. 6. (a) 256×256 pixel section of 1024×1024 pixel hologram of wedge-shaped sample of single-crystal MgO, and 180×150 pixel sections from 256×256 pixel reconstructed phase images with (b) no averaging, (c) 3×3 pixel averaging and (d) 10×10 pixel averaging. The white contour lines are plotted with $\pi/2$ increments in the phase.

detector caused a characteristic pattern of small intensity variations which, in the case of the Gatan 679, was a hexagonal pattern. These variations were taken out of the final image by gain normalization accomplished by on-line division by a gain-reference image [13].

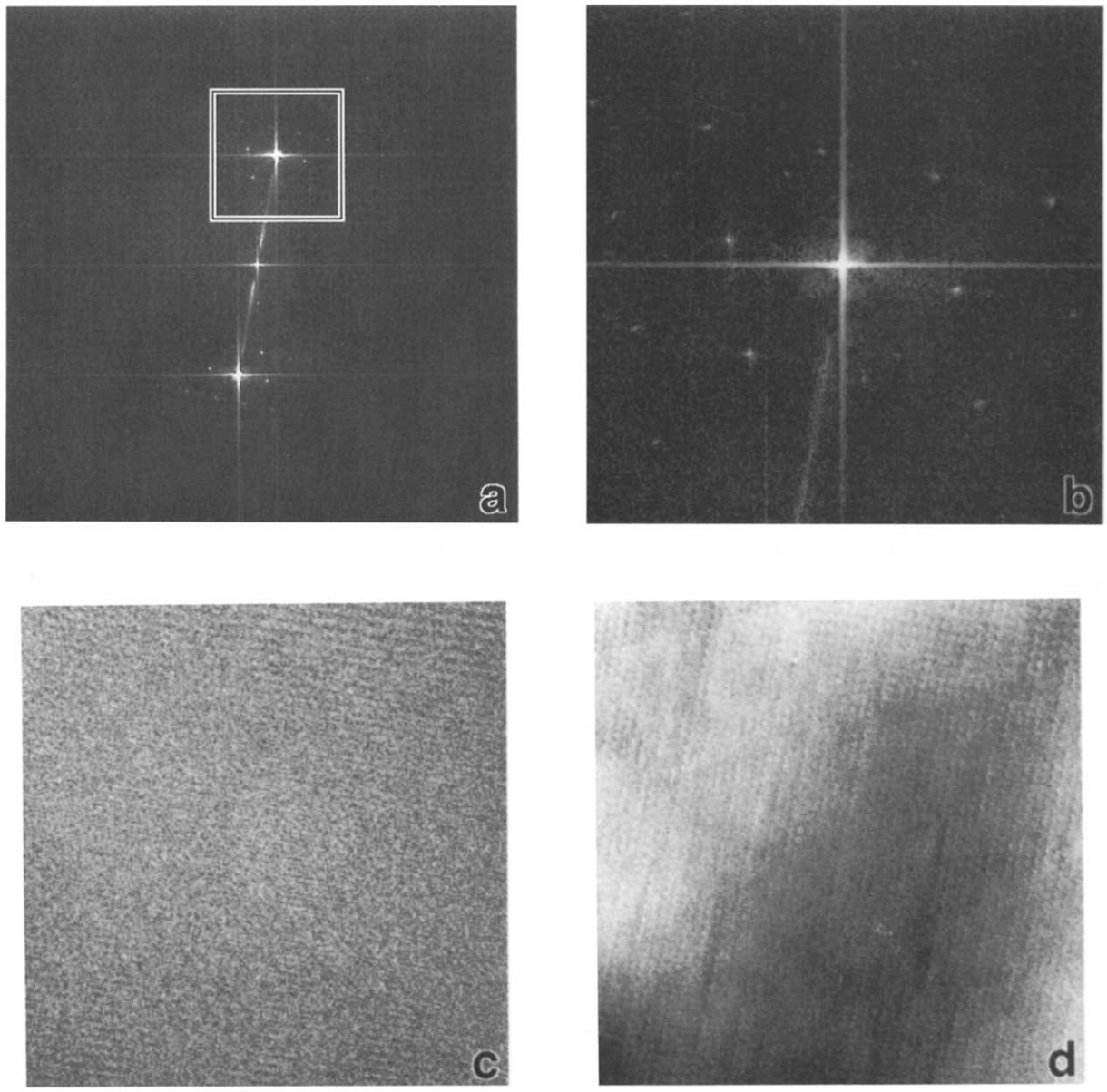

Fig. 7. (a) Fourier transform of a 1024×1024 reference hologram, (b) enlarged view of sideband in box indicated in (a), (c) reconstructed amplitude image, (d) reconstructed phase image, (e) intensity profile along a line in (d), (f) 30-line average centered along the same line, (g) intensity profile along a line in the corrected phase image and (h) 30-line average along the same line.

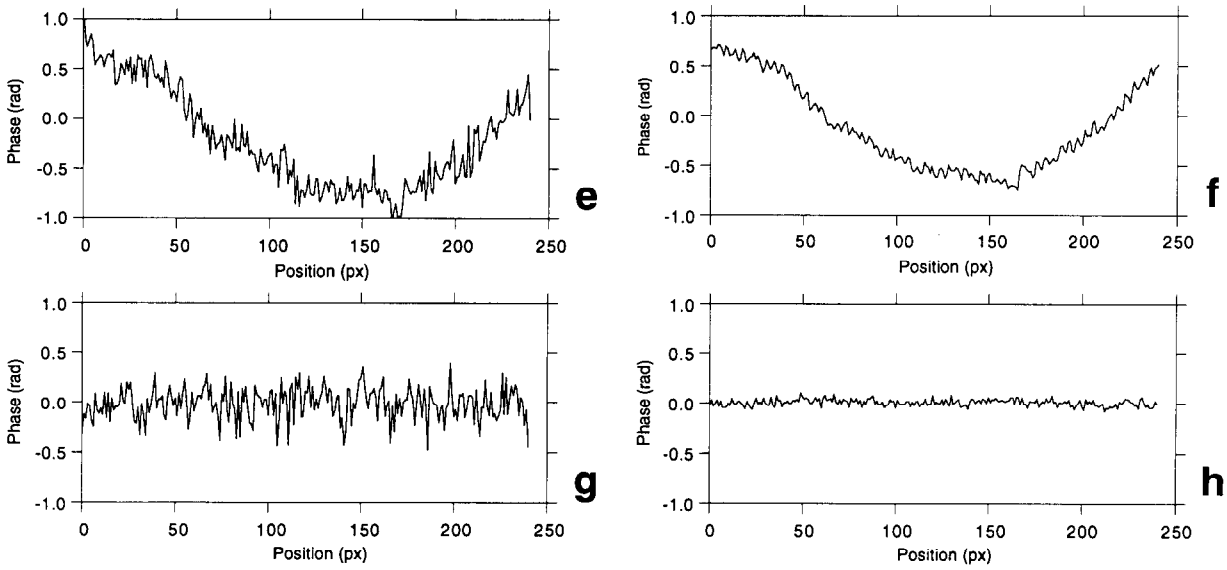

Fig. 7 (continued).

3.2. Phase reconstruction

The restoration of the phase from the digitally recorded hologram involved the use of standard Fourier processing techniques. The availability of digital data from the CCD camera greatly facilitated quantitative numerical processing and its precision, since there were no intermediate recording media or digitization processes which could distort the data.

A Fourier spectrum of the off-axis hologram shows prominent sidebands at a spatial frequency corresponding to the interference fringe spacing. The phase reconstruction involved the extraction of a subregion centered on one of the sidebands from the Fourier transform of the original hologram, as shown in fig. 5. Subsequent inverse Fourier transformation of this subregion yielded a complex image, from which the amplitude and phase were retrieved by simple mathematical operations.

Several artifacts of the imaging and reconstruction processes which appear at this point must be removed before quantification is attempted. First, geometric distortions of the projector lenses produce small shifts of the interference fringes which cause distortions of the reconstructed phase. Second, if the sideband is not chosen exactly at the center of the extracted subregion, a residual tilted plane will be added to the phase image. Finally, if the phase changes in the reconstructed image are larger than 2π, calculation of the phase by use of the inverse tangent causes artifacts in the reconstructed phase in the form of discontinuous jumps of 2π. Methods to remove these artifacts are discussed in the following sections. These methods have been used to produce the phase images in the example presented in fig. 6. The reconstructed phase images in figs. 6b–6d of surface steps present on a cleaved single-crystal wedge-shaped MgO sample clearly show that phase precision depends on the measurement area and that averaging techniques must be used to detect small phase differences.

3.3. Correction for geometric distortions and shear

Geometric distortion in an electron microscope image recorded on a CCD camera is usually too small to influence the image in any observable way. However, in off-axis electron holography, the phase of the wave at any position, if derived from the shift of an interference fringe, is very sensitive to distortions. Interference fringes

are typically recorded at spacings of about $4\Delta_c$ which, for the current generation of CCD arrays, corresponds to a maximum of about 100 μm. A distance of less than 100 μm at the detector then corresponds to a phase shift of 2π, which, due to the obtainable phase precision of about $\pi/100$, implies that distortions of the order of 1 μm at the detector can produce measurable phase shifts. Distortions of this magnitude were present in these experiments due to fiber optic shear and projector lens distortions.

The combined effect of both types of distortion was measured by recording a reference hologram (with no object present), since the reconstructed phase directly revealed the distortion pattern. This pattern was recorded and used to correct the phase of the image wave. This was easily accomplished by subtraction of the reconstructed distortion phase pattern from the reconstructed phase of the image wave.

A typical example of the Fourier transform of a 1024×1024 reference hologram is shown in fig. 7a. The typical spot pattern in the sidebands, clearly visible in the enlarged print in fig. 7b, is associated with fiber-optic shear. Fig. 7c shows the magnitude image indicating that gain-normalization of the hologram has effectively removed intensity variations due to camera fiber-optics, and fig. 7d shows the phase image representing a typical distortion pattern. The intensity profiles plotted in figs. 7e and 7f show a maximum local phase shift of about $\pi/2$, corresponding to a maximum local image shift of about 25 μm. We recorded reference holograms for every set of experiments performed at a given biprism voltage and lens settings. We found that the correction procedure was extremely effective and that the residual variations across the field of view attributable to geometric distortions were smaller than the shot noise, as can be seen in figs. 7g and 7h. Distortion correction in general doubles the variance of the noise in the reconstructed phase as can be seen from a comparison between figs. 7e and 7g. However, notice that the fluctuations present in the 30-line averaged corrected profile of fig. 7f are large than a similar profile in fig. 7h, indicating effective removal of the effects of fiber-optic shear by the distortion correction procedure. This leads to the important conclusion that distortion correction must always be performed, even when measuring phase differences on a small spatial scale, since fluctuations due to fiber-optic shear are in general larger than those caused by shot noise in the hologram.

3.4. Phase unwrapping

In order to properly analyze the phase of the image wave, it was necessary to calculate the unambiguously defined phase, which is commonly referred to as the "unwrapped phase". In most practical cases in electron holography, a simple phase-unwrapping algorithm based on detection of the 2π-discontinuities from the principal values of the phase is sufficient. We used an inverse tangent routine to determine the phase of the complex coefficients, and the discontinuities were removed by appropriately adding or subtracting 2π from areas of the phase image which were determined to be wrapped in phase. Detection of discontinuities was simply accomplished by monitoring the difference between two adjacent phase coefficients on a line-by-line basis. When the difference exceeded a pre-specified threshold, it was determined that a discontinuity was present. Before starting the line-by-line detection and unwrapping algorithm the starting phases were determined by analysis of the first column, as illustrated in fig. 8.

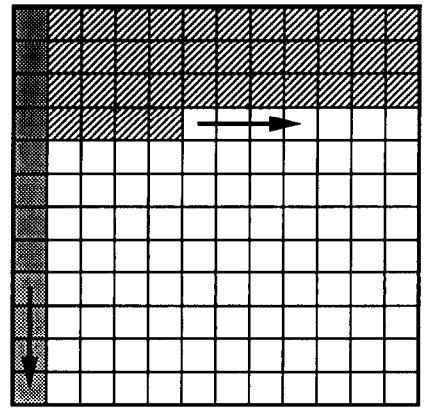

Fig. 8. Principle of phase unwrapping on a row-by-row basis. Before unwrapping the rows, the starting phases of the first column are determined.

This algorithm worked in most cases because the difference between adjacent samples of the unwrapped phase was always lower than the threshold. The main reason for failure was low signal-to-noise ratio. If the random phase deviations were larger than the wrapping detection threshold, it was no longer possible to unambiguously detect the discontinuities. This problem occurred when part of the sample was thick, producing low-amplitude interference fringes due to inelastic scattering, or when the interference pattern was recorded at very low dose.

The simplest method of phase unwrapping involved trying to minimize the phase variations in the image by choosing the center of the sideband (see section 3.5) and then shifting the phase in the complex image such that phase wrapping was avoided. The phase was shifted in the complex image by multiplying by a complex constant. Regions of the phase images at various phase shifts which were free of phase wrapping were then cut out and pasted together (after correcting for the relative phase shifts) to obtain an entire image which was free of phase wrapping. This method proved to be very effective and insensitive to noise, but was more tedious in practice than applying a single-pass unwrapping algorithm.

More sophisticated phase-unwrapping algorithms exist than those applied here. A one-dimensional algorithm which combines the information contained in the principal value and the derivative is given in ref. [26]. This approach could be extended to a two-dimensional scheme for improved performance.

3.5. Flattening the vacuum phase

A coordinate translation in reciprocal space by a vector $\boldsymbol{q}$ corresponds to a multiplication in real space by a plane wave $\psi(\boldsymbol{r}) = \exp(2\pi i \boldsymbol{q} \cdot \boldsymbol{r})$, which corresponds to an additional phase of $2\pi \boldsymbol{q} \cdot \boldsymbol{r}$ in the image. Since the extraction of the sideband from the Fourier transform of the hologram was generally only centered to within one pixel, the resulting reconstructed phase image also contained an additional phase resulting from the inaccurate centering. The form of the additional phase described above is a tilted plane, where the tilt will equal 2π over the width of the image for every pixel the extracted sideband is off-center.

Removal of this tilted plane was easily accomplished by fitting a plane of the form $I(x, y) = Ax + By + C$ by a least-squares method to an area of the phase image which showed only vacuum. The intensity in the plane was then subtracted over the entire image so that the phase was representative of that resulting from a plane wave normally incident on the sample.

4. Discussion

Detection limits in off-axis electron holography in the TEM are determined by the trade-off between precision and spatial resolution. Eq. (17) describes how the variance in the phase is inversely proportional to the measurement area. Since the dimensions of the measurement area A_R can be considered to correspond to the spatial resolution, eq. (17) implies that the precision of the phase measurement is inversely proportional to the spatial resolution. For low spatial resolution, say details larger than 2–5 nm, the measurement precision can be improved by measuring the phase averaged over a large region. However, if effects must be studied on a short spatial scale, say 0.5–1 nm, considerable deviations in the phase due to shot noise must be expected. Therefore, for measurement of small phase differences, such as surface steps in thin films, the typical resolution in practice will be about 2–10 nm (see fig. 6). If large phase differences occur, for example at certain heterogeneous interfaces, larger phase errors can be tolerated and materials properties can be studied with spatial resolutions as low as 0.5–1 nm [14].

Implicit in all of the previous discussions is the fact that the shot noise which limits the measurement precision is dependent entirely on the number of electrons collected from the region in which the phase is to be measured. Therefore, the important parameter is the area of the region of interest, with no mention of the shape of that area. This independence of the noise from the aspect ratio of the region can be exploited in specimens which contain only one-dimensional

variations of potential. By averaging in the direction in which there are no spatial variations, the signal-to-noise ratio can be increased without degrading the spatial resolution in the perpendicular direction.

It should be considered what limits the ultimate spatial resolution. For spatial frequencies above about 2 nm^{-1}, the objective lens introduces significant mixing of phase and amplitude information, where the magnitude of the effect is dependent on the spatial frequency. Therefore, in order to extract any information at a resolution better than about 0.5 nm, it would be necessary to carry out a full correction of the objective lens aberrations [8], although it is possible to minimize this mixing effect by choosing the proper defocus.

The application of eqs. (15)–(17) to experimental design also needs to be discussed, in order to evaluate the influence of the adjustable experimental parameters on the precision of the experimental measurement. The parameters associated with the gun depend largely on the type of gun (thermionic or field emission), and, since the product $\beta R_s \alpha_0$ is at least 100 times larger for field emission guns than for thermionic guns, field emission guns are obviously the electron source of choice for electron holography. The objective lens parameters M_{obj} and f are relatively fixed and do not vary greatly over different microscopes, although the new low-gap objective lenses currently being implemented could offer an improvement in the phase precision by at least a factor of 2. Obviously, the biprism distances a and b are variable (especially b), but, as with the objective lens parameters, improvement of the phase precision by adjustment of these parameters is accompanied by a necessary increase in the image magnification (leading to a reduced field of view) and the biprism voltage.

Eqs. (15)–(17) imply that the number of pixels on the CCD array should be kept to a minimum in order to keep the image magnification and biprism voltage down and to reduce the phase uncertainty. This interesting result arises from the rather restrictive requirement in eq. (13) that the whole interference pattern fit on the detector. This requirement implies that a fixed current (depending only on the electron gun parameters) would be incident on the detector which should all be focused in the smallest possible number of fringes (where the number of fringes should be N/n). This is an impractical requirement, so the number of pixels in the CCD array should be considered a constant.

5. Conclusions

Application of slow-scan CCD cameras for digital acquisition of electron holograms has permitted quantitative measurement of the phase of the image wave with a precision higher than that obtained from holograms recorded on conventional photographic plates. In addition, reconstruction and analysis from digital holograms are much more straightforward than from photographic plates, since the tedious optical reconstruction and linearization techniques are now avoided.

Precise measurement of the phase of the exit-surface wave requires three correction procedures, which have been implemented with digital image processing. First, it has been established that geometric distortions of the projector lens system can be corrected within residual phase errors of $\pi/100$ using a reference hologram recorded in the absence of a specimen. Second, phase unwrapping is required and we have shown that in most cases a simple phase algorithm, based on detection of discontinuities in the primary value of the phase, is sufficient to obtain the correct phases. Finally, absolute phase measurements can be obtained by using the phase in a region showing vacuum as a reference.

Electron holograms are typically recorded with low electron dose (100–200 el/px) which can lead to phase measurements with significant errors due to shot noise. Errors can be minimized by sensible choices of the excitation of the first intermediate lens, image magnification, biprism voltage, exposure time and the use of astigmatic illumination with optimized diameters of the major and minor axes. Further reduction of statistical errors can be accomplished by averaging techniques, however, in general at the expense of spatial resolution.

Acknowledgements

We thank Dr. M. Gajdardziska-Josifovska for permission to publish results from her work on MgO surface steps. We gratefully acknowledge Drs. H. Lichte and E. Völkl for their help in realizing holography on our EM400ST-FEG. Thanks are due to Drs. M.R. McCartney and David J. Smith for their support and many useful discussions. This work was performed at the Center for High-Resolution Electron Microscopy within the Center for Solid State Science at Arizona State University, supported by NSF grants DMR 89-13384 and 91-15680. Support for J.K.W. was also provided by the Industrial Associates Program at Arizona State University.

Appendix A. Asymptotic Cramér-Rao lower bound for the parameters of holography fringes

This appendix presents the minimum variance bound for estimation of the parameters (spatial frequency, amplitude and phase) of holography fringes. The observed hologram intensity $\underline{w}(m, n)$, where m and n are the pixel coordinates (underlined characters denote stochastic variables), can be written as:

$$\underline{w}(m, n) = B + A \sin(mu + nv + \phi) + \underline{\epsilon}(m, n), \tag{A.1}$$

where $\underline{\epsilon}(m, n)$ models the shot-noise contribution (which is Poisson-distributed and white), B is the background intensity, A and ϕ denote the amplitude and the phase of the holography fringes, respectively, and u and v are the components of the spatial frequency vector of the holography fringes.

The calculations in this appendix are based on previously published results [27] valid for one-dimensional signals. Basic considerations and common notation can be found in standard textbooks on statistical parameter estimation, for example ref. [28]. The Cramér-Rao lower bound, which defines the lower bound for the variance of any unbiased estimator for the holography-fringe parameters, is defined by the inequality $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}_{\mathrm{CR}}\boldsymbol{x} \geq \boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{x}$ valid for any vector $\boldsymbol{x}$ (T denotes transposition), where $\boldsymbol{P}_{\mathrm{CR}}$ denotes the variance-covariance matrix $\boldsymbol{P}_{\mathrm{CR}} = \mathrm{cov}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})$, with $\hat{\boldsymbol{\theta}}$ being an estimator for the parameter vector $\boldsymbol{\theta}$ given by $\boldsymbol{\theta} = (A, \phi, u, v)^{\mathrm{T}}$ and $\mathrm{cov}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})$ being defined by its i, j element $\mathrm{cov}(\hat{\theta}_i, \hat{\theta}_j)$. $\boldsymbol{M}$ is the Fisher information matrix given by

$$\boldsymbol{M} = E\left[\left(\frac{\partial \ln[f(\underline{\boldsymbol{w}};\boldsymbol{\theta})]}{\partial\boldsymbol{\theta}}\right)^{T}\left(\frac{\partial \ln[f(\underline{\boldsymbol{w}};\boldsymbol{\theta})]}{\partial\boldsymbol{\theta}}\right)\right], \tag{A.2}$$

where E denotes the expectation operator and $f(\underline{\boldsymbol{w}};\boldsymbol{\theta})$ is the joint probability density of the pixel intensities of the observed hologram $\underline{w}(m, n)$, with $\underline{\boldsymbol{w}}$ an one-dimensional vector containing the values of $\underline{w}(m, n)$. If the disturbances $\underline{\epsilon}(m, n)$ are independent and identically distributed Gaussian random variables with variance σ^2, which is a valid assumption if the mean intensity in the hologram is not too small and the fringe visibility is not too large, then the joint probability density of the observations is

$$f(\underline{\boldsymbol{w}};\boldsymbol{\theta}) = \prod_{m=0}^{M-1}\prod_{n=0}^{N-1}\frac{1}{\sigma\sqrt{2\pi}} \times \exp\left[-\frac{1}{2\sigma^2}\underline{d}^2(m, n;\boldsymbol{\theta})\right], \tag{A.3}$$

with

$$\underline{d}(m, n;\boldsymbol{\theta}) = \underline{w}(m, n) - A \sin(mu + nv + \phi). \tag{A.4}$$

It is noted that the background intensity B and the variance of the noise σ^2 could also be included as unknown parameters. However, for one-dimensional signals it has already been established that these parameters do not influence the precision of estimators for fringe parameters [27]. This result also holds for two dimensions [29], and therefore B and σ^2 are not included here. The elements of the information matrix (A.2) follow from basic calculations, and are given by:

$$\boldsymbol{M} = \frac{1}{\sigma^2}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\frac{\partial\underline{d}(m, n;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\left(\frac{\partial\underline{d}(m, n;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^{T}. \tag{A.5}$$

It is difficult to derive an exact, closed-form solution for eq. (A.5) valid for any M and N. However, it has been shown that it is possible to obtain a simple expression for the asymptotic Cramér-Rao lower bound [27]:

$$P_{CR}^{\infty} = \lim_{M,N\to\infty} M^{-1}.$$

P_{CR}^{∞} can be calculated using:

$$\lim_{M,N\to\infty} \frac{1}{M^{1+q}} \frac{1}{N^{1+r}} \times \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} m^q n^r \cos(mu + nv + \phi) = \begin{cases} \frac{1}{1+q} \frac{1}{1+r} \cos\phi & \text{if } u = 0 \text{ and } v = 0, \\ 0 & \text{if } u \neq 0 \text{ or } v \neq 0, \end{cases} \tag{A.6}$$

which is valid for $q \geq 0$, $r \geq 0$ and $-\pi \leq u < \pi$, $-\pi \leq v < \pi$. Eq. (A.6) directly follows from the one-dimensional equivalent presented in ref. [27]. Element 1,1 of M follows from:

$$\begin{aligned} &\lim_{M,N\to\infty} \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \frac{\partial \underline{d}(m, n; \boldsymbol{\theta})}{\partial A} \frac{\partial \underline{d}(m, n; \boldsymbol{\theta})}{\partial A} \\ &= \lim_{M,N\to\infty} \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sin[mu + nv + \phi] \times \sin[mu + nv + \phi] \\ &= \frac{1}{2} \lim_{M,N\to\infty} \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \cos[0] - \cos[2(mu + nv + \phi)] = \tfrac{1}{2}, \end{aligned} \tag{A.7}$$

where the derivatives in eq. (A.7) are calculated from eq. (A.4). Comparison of eq. (A.7) with eq. (A.5) implies that element 1,1 of M equals $MN/2\sigma^2$. All other elements of M follow in similar fashion from eq. (A.5), by repeated use of eq. (A.6) with appropriate values for q and r, resulting in the asymptotic Cramér-Rao lower bound:

$$P_{CR}^{\infty} = \begin{bmatrix} \frac{MN}{2\sigma^2} & 0 & 0 & 0 \\ 0 & \frac{MNA^2}{2\sigma^2} & \frac{M^2NA^2}{4\sigma^2} & \frac{MN^2A^2}{4\sigma^2} \\ 0 & \frac{M^2NA^2}{4\sigma^2} & \frac{M^3NA^2}{6\sigma^2} & \frac{M^2N^2A^2}{8\sigma^2} \\ 0 & \frac{MN^2A^2}{4\sigma^2} & \frac{M^2N^2A^2}{8\sigma^2} & \frac{MN^3A^2}{6\sigma^2} \end{bmatrix}^{-1}. \tag{A.8}$$

Inversion of the matrix yields:

$$P_{CR}^{\infty} = \begin{bmatrix} \frac{2\sigma^2}{MN} & 0 & 0 & 0 \\ 0 & \frac{14\sigma^2}{MNA^2} & -\frac{12\sigma^2}{M^2NA^2} & -\frac{12\sigma^2}{MN^2A^2} \\ 0 & -\frac{12\sigma^2}{M^2NA^2} & \frac{24\sigma^2}{M^3NA^2} & 0 \\ 0 & -\frac{12\sigma^2}{MN^2A^2} & 0 & \frac{24\sigma^2}{MN^3A^2} \end{bmatrix}. \tag{A.9}$$

The matrix elements on the diagonal of eq. (A.9) give expressions for the minimum variance bounds for estimation of the amplitude, the phase and the spatial frequency in the x and the y direction, respectively. Non-diagonal elements give expressions for the covariances between these parameters.

In this paper, we are interested in the minimum variance bound for the phase expressed in terms of visibility of the holography fringes and electron dose. This bound can be easily derived from eq. (A.9) and the following considerations. If the number of electrons per pixel is N_e, then the total dose N_t in the measurement area of $M \times N$ pixels is $N_t = MNN_e$, the variance of the shot noise is $\sigma^2 = N_e$, and the amplitude of the holography fringes is $A = VN_e$ (V denotes visibility). The 2,2 element of eq. (A.9) implies that the minimum variance bound of the estimate of the phase is:

$$\mathrm{var}[\hat{\phi}] = \frac{1}{MN} \frac{14\sigma^2}{A^2} = \frac{14}{V^2 N_t}. \tag{A.10}$$

The variance bound (A.10) is an asymptotic expression true for a measurement area with a large number of pixels, that is derived using the assumption that the noise is Gaussian-distributed and stationary up to second order (in other words, the standard deviation is spatially invariant). This assumption of stationary noise is only approximately true. However, simulation experiments [29] indicated that eq. (A.10) still holds under practical conditions where MN can be as small as 16×16 pixels. For extremely small measurement regions ($< 6 \times 6$ pixels) this asymptotic expres-

sion no longer holds and it turns out that the value of the phase influences its measurement precision considerably [11].

Notice, that if the spatial frequency in the x and the y direction are a-priori known, the terms associated with u and v should be excluded from eq. (A.8) and the inversion of the resulting 2×2 matrix gives $\mathrm{var}[\hat{\phi}] = 2/V^2 N_t$, in agreement with refs. [17,18].

References

[1] A. Tonomura, Rev. Mod. Phys. 59 (1987) 639.
[2] A. Tonomura, Electron holography, in: Progress in Optics 23 (North-Holland, Amsterdam, 1986) p. 183.
[3] S. Frabboni, G. Matteucci, G. Pozzi and M. Vanzi, Phys. Rev. Lett. 55 (1985) 2196.
[4] G. Möllenstedt and H. Düker, Naturwissenschaften 42 (1955) 41.
[5] C. Jonsson, H. Hoffmann and G. Möllenstedt, Phys. Kondens. Mater. 3 (1965) 193.
[6] K.-J. Hanszen, Adv. Electron. Electron Phys. 59 (1982) 1.
[7] K.-J. Hanszen, J. Phys. D (Appl. Phys.) 19 (1986) 373.
[8] H. Lichte, Ultramicroscopy 20 (1986) 293.
[9] S. Hasegawa, T. Kawasaki, J. Endo, A. Tonomura, Y. Honda, M. Futamoto, K. Yoshida, F. Kugiya and M. Koizumi, J. Appl. Phys. 65 (1989) 2000.
[10] G. Ade and R. Lauer, Optik 91 (1992) 5.
[11] E. Völkl, High Resolution Electron Holography, PhD Thesis, Eberhard-Karls-Universität, Tübingen, 1991.
[12] P.E. Mooney, G.Y. Fan, C.E. Meyer, K.V. Truong, D.B. Bui and O.L. Krivanek, in: Proc. 12th Int. Congr. on Electron Microscopy, Seattle, WA, 1990, Vol. 1 (San Francisco Press, San Francisco, CA, 990) p. 164.
[13] W.J. de Ruijter and J.K. Weiss, Rev. Sci. Instr. 63 (1992) 4314.
[14] J.K. Weiss, W.J. de Ruijter, M. Gajdardziska-Josifovska, M.R. McCartney and D.J. Smith, Ultramicroscopy 50 (1993) 301.
[15] M. Gajdardziska-Josifovska, M.R. McCartney, W.J. de Ruijter, D.J. Smith, J.K. Weiss and J.M. Zuo, Ultramicroscopy 50 (1993) 285.
[16] M. Born and E. Wolf, Principles of Optics (Pergamon, Oxford, 1980) p. 505.
[17] H. Lichte, K.-H. Herrmann and F. Lenz, Optik 77 (1987) 135.
[18] J.F. Walkup and J.W. Goodman, J. Opt. Soc. Am. 63 (1973) 399.
[19] W.J. de Ruijter, P.E. Mooney and O.L. Krivanek, in: Proc. 51th Annual MSA Meeting, 1993, accepted for publication.
[20] V.V. Fedorov, Theory of Optimal Experiments (Academic Press, New York, 1972).
[21] A.V. Crewe, in: Progress in Optics, Vol. 11 (North-Holland, Amsterdam, 1973) p. 225.
[22] P. Pozzi, Optik 77 (1987) 69.
[23] F.F. Medina and G. Pozzi, J. Opt. Soc. Am. A 7 (1990) 1027.
[24] J.K. Weiss, R.W. Carpenter and A.A. Higgs, Ultramicroscopy 36 (1991) 391.
[25] J.K. Weiss, W.J. de Ruijter, M. Gajdardziska-Josifovska, D.J. Smith, E. Völkl and H. Lichte, in: Proc. 49th Annual EMSA Meeting, San Jose, CA, 1991 (San Francisco Press, San Francisco, CA, 1991) p. 674.
[26] J.M. Tribolet, IEEE Trans. Acoust. Speech Signal Process. 25 (1977) 170.
[27] P. Stoica, R.L. Moses, B. Friedlander and T. Soderstrom, IEEE Trans. Acoust. Speech Signal Process. 37 (1989) 378.
[28] A. van den Bos, Parameter estimation, in: Handbook for Measurement Science (Wiley, New York, 1982).
[29] W.J. de Ruijter, Quantitative High-Resolution Electron Microscopy and Holography, PhD Thesis (Delft University Press, Delft, 1992).