Scanning Microscopy Vol. 11, 1997 (Pages 437-454) 0891-7035/97$5.00+.25
Scanning Microscopy International, Chicago (AMF O'Hare), IL 60666 USA

THE USE OF STOCHASTIC ALGORITHMS FOR PHASE RETRIEVAL IN HIGH RESOLUTION TRANSMISSION ELECTRON MICROSCOPY

A. Thust*, M. Lentzen and K. Urban

Institut für Festkörperforschung, Forschungszentrum Jülich GmbH, Jülich, Germany

Abstract

The usefulness of stochastic algorithms to retrieve the exit-plane wave function from periodic high-resolution electron microscopic images is investigated. In contrast to "classical" focal series reconstruction methods which need approximately twenty input images, fully non-linear reconstructions of periodic wave functions are possible in most cases from only two images using stochastic algorithms. The efficiency and accuracy of two different algorithms, a Simulated Annealing algorithm and a genetic algorithm, are compared with each other. Simulated high-resolution images of the intermetallic alloy Ni_4Mo and experimental images of the high-T_c superconductor $YBa_2Cu_3O_7$ are used as a test input.

Key Words: High-resolution transmission electron microscopy, phase retrieval, reconstruction, exit-plane wave function, simulated annealing, genetic algorithms, stochastic algorithms.

*Address for correspondence:
A. Thust
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH
D-52425 Jülich, Germany
Telephone number: ++49 2461 616644
FAX number: ++49 2461 616444
E-mail: A.Thust@fz-juelich.de

Introduction

During the past twenty years, considerable effort has been put into the development of techniques for the retrieval of the exit-plane wave function (EPW) in high-resolution electron microscopy. These techniques have proven to be of considerable advantage, especially when making use of the information lying beyond the Scherzer resolution limit of the microscope. The high-resolution information stemming from the frequency band between Scherzer limit and information limit is often not suitable for an image interpretation by eye. Due to the spherical aberration and the defocusing of the objective lens, the electron wave emerging from the object is distorted when propagating to the final recording plane. These distortions become considerably stronger with increasing resolution. It is therefore advantageous to retrieve the EPW, since it is unaffected by imaging artifacts up to the information limit of the microscope.

Besides electron holography, the technique of focus variation is a well-known approach for the retrieval of the EPW in transmission electron microscopy. The amplitude and phase of the EPW are retrieved by making use of the information contained in a series of images taken from the same object area but with a different objective lens defocus for each single image. Prominent algorithmic approaches to solving the phase retrieval problem in focus-variation microscopy are the so-called Paraboloid Method (see, e.g., Schiske, 1973; Saxton, 1978, 1980, 1986, 1993, 1994; Van Dyck and Op de Beeck, 1990, 1993) and the Maximum Likelihood method (see, e.g., Kirkland, 1984; Kirkland *et al*., 1985; Coene *et al*., 1992, 1996; Thust *et al*., 1996a,b).

The Paraboloid Method (PAM) aims at a separation of the EPW from its complex conjugate counterpart and at an extraction of linear contrast contributions from the total image contrast. Using N input images, both effects improve roughly by a factor $\sqrt{N}$ in the final wave function. Useful results of the PAM over a sufficiently large interval of spatial frequencies can therefore be only expected when taking at least 15 to 20 images as input. A second restriction of the PAM is due to the fact that it concentrates on the linear contrast contributions which may be insufficiently separated from the nonlinear contributions in the case of strongly

scattering objects. Recursive implementations of the PAM which compensate for strong nonlinear contributions to the image contrast turned out to be of limited use (Thust *et al.*, 1996a).

A more powerful algorithm to retrieve the EPW from a focal series is the Maximum Likelihood (MAL) method. The MAL algorithm is based on a recursive feedback principle which exploits the difference between simulated images based on a trial EPW and experimental through-focus images. By minimizing the difference between simulated and experimental images, one assumes that the trial EPW converges towards the actual experimental EPW. Contrary to the PAM, which aims at a complete elimination of nonlinear contrast contributions, the MAL method exploits actively nonlinear contrast phenomena. As a consequence, the MAL method can also be employed for strongly scattering objects. As is the case for the PAM, the MAL method requires a set of typically 20 input images. The use of a substantially lower number of input images is in most cases not possible. Due to the strong nonlinear coupling of diffracted beams, a local minimization principle, as is the least squares formalism applied in the MAL method, may fail.

At this point, a question arises: how many images are in principle required for a successful reconstruction? The reproducible acquisition of 20 or more images needed for the PAM or MAL method is still not a trivial experimental task nowadays. Apart from specimen drift, the sample might suffer from radiation damage or may bend due to the heating effect of the electron beam. It is therefore beneficial to have the capability to perform reconstructions of the EPW using as few as possible input images. It was shown that two micrographs taken under different defocus can provide sufficient information to reconstruct the EPW of a strongly scattering object (Drenth *et al.*, 1975). A well-known iterative procedure based on two input images (Misell, 1973) does, however, not take into account the resolution limiting effect of partially coherent illumination (see, e.g., Frank, 1973).

In the present paper, we investigate an alternative approach for the fully nonlinear reconstruction of the EPW from through-focus images which is mainly suitable for periodic objects. This approach, which is based on the use of stochastic algorithms, requires a substantially smaller number of input images than the well-established PAM or MAL methods.

Theory

Categorization of reconstruction algorithms

The most efficient approach for phase retrieval would be to find an analytical expression which inverts the imaging process and transforms the input images directly back into the EPW. Such an analytical inversion has not been achieved due to the non-linearity of the imaging process. A solution close to the analytical approach is the Paraboloid Method (PAM) in the case that the linear approximation to the image formation holds. However, the nonlinear interference terms which become important for thicker objects cannot be treated in such a straightforward way.

A more tedious approach in terms of numerical effort has to be chosen in the case of strongly nonlinear image formation. The Maximum Likelihood (MAL) approach is based on the optimization of an initial guess to the EPW. The images calculated from this trial EPW are compared with the experimental images, and a feedback to the EPW is calculated on the basis of this comparison. One assumes that the correct EPW is found if the squared intensity difference between input images and calculated images is minimized. In the following, the quantity to be minimized will be more generally called evaluation function, since it expresses the ability of a particular trial EPW to reproduce the experimentally observed contrast. Within the MAL formalism, the minimization of the evaluation function requires multiple feedback (backward) and image calculation (forward) steps instead of only one backward step needed within the PAM formalism. Convergence to the correct solution is only guaranteed if one unique minimum of the evaluation function exists. The existence of only one unique minimum is difficult to prove and depends also on the number of input images given to the algorithm. The less images are used, the more it is likely that the evaluation function exhibits local minima which may cause a failure of the algorithm due to its exclusively local search characteristics.

If the search space is covered densely by local minima and maxima of the evaluation function, a global search strategy must be applied. The most extreme form of a global search strategy is random search. By trying out randomly different EPWs, one has in principle a chance to find the correct EPW which reproduces the experimentally observed focal series. It is, however, easy to demonstrate that such a strategy would exceed the capacity of the fastest computers and is not viable for the solution of practical problems.

From the above considerations, it can be concluded that a compromise between local search and random search should be most efficient in cases where local minima are encountered, as is the case when attempting nonlinear reconstructions from only a few input images. Compared to the PAM formalism and to the more calculation intensive MAL formalism, the computational load should again increase strongly due to the need for a global search strategy.

During the past few decades, two highly efficient types of algorithms have been developed which meet the demand to combine global and local search strategies. Genetic algorithms, as well as the technique of Simulated Annealing, have proven to be valuable tools for finding the best solution for problems exhibiting local optima. A good

overview over these algorithms can be found in the work of Davis and Steenstrup (1987). In the framework of focal-series reconstruction, these stochastic algorithms can be roughly positioned between the MAL algorithm and the impractical random search approach because the search characteristics can be adjusted continuously between pure local search and pure random search, depending on the particular demands.

In order to keep the computational effort within reasonable limits, the present application of stochastic algorithms is limited to periodic objects. Typical periodic wave functions in high-resolution electron microscopy can be decomposed into a small number of Fourier coefficients which does in most cases not exceed one hundred. The high demand in global search activity is compensated in this work by a restriction to periodic objects keeping the application of stochastic algorithms still well suited for practical use.

The evaluation function

In the framework of stochastic algorithms, it is necessary to rate frequently a set of input parameters with respect to the desired output. In the present context, the input parameters describe a particular trial EPW, and the output to be optimized is the ability of this EPW to reproduce the experimentally observed image contrast. Two basic procedures are therefore required to evaluate the EPW: firstly, a procedure to simulate a focal series of images based on the EPW and, secondly, a figure-of-merit describing the match between the simulated images and the experimental images.

Image calculation

In the following, the exit-plane wave function $\Phi(\vec{r})$ is denoted in its Fourier-space representation, where the Fourier coefficients $F(\vec{k})$ are given by

$$F(\vec{k}) = \int \Psi(\vec{r}) e^{2\pi i \vec{k}\vec{r}} \, d\vec{r} \tag{1}$$

For partially coherent illumination, the Fourier-space description of the resulting image intensity $I(\vec{k})$ is given by (Frank, 1973):

$$I(\vec{k}) = \sum_{\vec{g}} F(\vec{g}+\vec{k}) F^*(\vec{g}) T(\vec{g}+\vec{k}, \vec{g}, Z) \tag{2}$$

Equation (2) is a weighted autocorrelation in Fourier space with the complex transmission cross-coefficient (TCC) T as a weighting factor. Renaming the argument vectors $\vec{a}$ and $\vec{b}$, the transmission cross-coefficient $T(\vec{a},\vec{b},Z)$ describes the phase shifts and damping effects imposed by the microscope on the mutual interference terms between two beams with wave vectors $\vec{a}$ and $\vec{b}$ at a defocus value Z (Ishizuka, 1980). Linear imaging theory takes only such terms $F(\vec{a})\cdot F^*(\vec{b})\cdot T(\vec{a},\vec{b},Z)$ into account where $|\vec{a}| = 0$ or $|\vec{b}| = 0$. This condition means that only the interference terms between diffracted beams and the unscattered beam are evaluated.

Given a wave function consisting of N beams, a complete nonlinear image simulation following the TCC formalism requires the calculation of N^2 interference terms. From a computational point of view, a dramatically faster alternative can be used for the highly coherent illumination produced by field emission guns (FEGs) (Coene *et al*., 1992, 1996). Nevertheless, the TCC formalism can still be employed very efficiently in the context of Simulated Annealing. During the nth cycle of the algorithm, it is not necessary to calculate the intensity coefficients $I_n(\vec{k})$ based on a certain wave function Ψ_n completely from scratch, but it is only necessary to make a small update to the already known intensity coefficients $I_{n-1}(\vec{k})$ based on a previous wave function Ψ_{n-1}. This is possible due to the fact that from one iteration step to the next only one single Fourier coefficient $F(\vec{g})$ of the previous wave function is changed within the Simulated Annealing algorithm, and only intensity differences $\Delta I(\vec{k})$ due to this change have to be actually calculated. All Fourier coefficients of the image intensity $I_n(\vec{k})$ belonging to the new wave function Ψ_n are related to the previous coefficients $I_{n-1}(\vec{k})$ via

$$I_n(\vec{k}) = I_{n-1}(\vec{k}) + \Delta I(\vec{k}) \tag{3}$$

If only a single Fourier coefficient $F(\vec{g})$ of the wave function is changed by the algorithm, the difference in image intensity is given for $\vec{k} \neq 0$ by

$$\begin{aligned}\Delta I(\vec{k}) = {} & \Delta F(\vec{g}) F^*(\vec{g}-\vec{k}) T(\vec{g},\vec{g}-\vec{k},Z) + \\ & F(\vec{g}+\vec{k}) \Delta F^*(\vec{g}) T(\vec{g}+\vec{k},\vec{g},Z)\end{aligned} \tag{4}$$

where the change of the wave function related to the wave vector $\vec{g}$ has been denoted as

$$\Delta F(\vec{g}) = F_n(\vec{g}) - F_{n-1}(\vec{g}) \tag{5}$$

Compared to the complete formulation given in Equation (2), the updating technique requires substantially less computational effort. Whereas the complete formulation described by Equation (2) involves the calculation of N^2 interference terms, the updating technique described by Equation (4) allows reduction of the number of interference terms to $2N$-1. For the sake of simplicity, the possibility of exploiting the Friedel symmetry of the Fourier transform of the image intensity has not yet been considered. If the Friedel symmetry is additionally exploited, only N interference terms have actually to be taken into account for the fully nonlinear update, and the numerical effort is thus the same as that needed for a linear image calculation.

Image comparison

The ability of a trial EPW to reproduce the experimental image contrast is assessed by means of a quality

factor. This quality factor expresses the match between the experimental images and the images calculated on the basis of a certain wave function Ψ. From an algorithmic point of view, the choice of a particular evaluation function is free. Any of the various measures used commonly for image comparison can be used. One can use measures like the squared difference intensity, a χ^2 measure, or a measure based on the cross-correlation between images. Since, in all cases, the information content of two images is projected onto one number, none of the possible measures can be regarded as perfect or superior to all others. Due to the projection onto one number, the discrepancies between two images are weighted differently by each measure, and the particular choice of a measure depends on which intensity differences should be stressed and which should be regarded as less significant.

We use a measure based on the cross-covariance coefficient for the evaluation function of our stochastic algorithms. This measure neglects differences between the image mean values as well as differences of the absolute contrast scaling of two images. The feature compared by this measure is the pattern content which turned out to provide sufficient information for the retrieval of the EPW.

It is advantageous to perform the image comparison in Fourier space for two reasons. First, the preceding image simulation step is carried out in Fourier space and a transform to real space can be avoided. Secondly, since we deal here with periodic objects, the Fourier transform consists only of relatively few coefficients, and image comparison can be accelerated considerably compared to a real-space procedure.

The simulated and the experimental image transforms can be arranged in vector form, where we use the symbol $\vec{e}$ for the experimental and the symbol $\vec{s}$ for the simulated image vector. The dimension M of the vectors corresponds to the number of involved Fourier coefficients. Denoting the Fourier coefficients of the experimental image with $I_e(\vec{k})$ and those of the simulated image with $I_s(\vec{k})$, the vectors $\vec{e}$ and $\vec{s}$ are written as

$$\vec{e} = (I_e(\vec{k}_l),...,I_e(\vec{k}_M)) \tag{6a}$$

$$\vec{s} = (I_s(\vec{k}_l),...,I_s(\vec{k}_M)) \tag{6b}$$

where the image mean values $I_{e,s}(\vec{k} = 0)$ have to be excluded in order to obtain the cross-covariance instead of the cross-correlation coefficient. With this notation, the normalized cross-covariance coefficient c can be displayed in the compact form

$$c = \frac{\vec{e} \cdot \vec{s}^*}{|e||s|} \tag{7}$$

where the asterisk denotes the complex conjugate, and the notations $|\vec{e}|$ and $|\vec{s}|$ are short forms for the image contrast given by $(\vec{e}\cdot\vec{e}^*)^{1/2}$ and $(\vec{s}\cdot\vec{s}^*)^{1/2}$, respectively. The coefficient c of Equation (7) is the cosine of the angle φ between the vectors $\vec{e}$ and $\vec{s}$. We use the angle φ instead of the coefficient c as a measure describing the ability of a wave function Ψ to reproduce the experimentally observed image contrast. For a focal series consisting of N images taken at different defocus values Z_i, we define the evaluation function $E(\Psi)$ as the mean angle $\langle\varphi\rangle$ with

$$E(\Psi) = <\varphi> = \frac{1}{N}\sum_{i=1}^{N}\arccos(c(Z_i)) \tag{8}$$

If $E(\Psi)$ amounts to 90° the images simulated on the basis of the wave function Ψ are completely uncorrelated with those of the experimental series. In contrast, $E(\Psi) = 0^\circ$ indicates a perfect correspondence between the respective image patterns.

Numerical efficiency

The search space to be investigated for an N-beam reconstruction may increase exponentially with the number of nonlinearly coupled beams. As a consequence, a similar increase of the required calls to the evaluation function is expected. Since almost the complete numerical effort is spent for frequent calculations of the evaluation function, it is important to keep the computation time related to the evaluation function as short as possible.

The calculation of the evaluation function is a two-step process. First, the images based on a trial wave function are calculated and, second, these simulated images are compared with the experimental images (see previous two sections). In this context, it is interesting to estimate the final gain in numerical efficiency that can be achieved by employing the updating technique for the image calculation instead of the image calculation from scratch.

In the following, the possibility to exploit the Friedel symmetry of the image intensity will be ignored, since it cancels out in the final result. The computation time required for the calculation of the evaluation function can be roughly related to the required number of complex multiplications. The number of complex multiplications needed for one image calculation from scratch is $2N^2$ (Eqn. 2), whereas the number of multiplications involved in the updating technique is close to $4N$ (Eqn. 4). The image comparison step is identical for both alternatives and requires approximately $8N$ complex multiplications. The estimate of $8N$ multiplications is based on two considerations: first, due to the nonlinear image formation, the Fourier transform of the image extends twice as far in Fourier space than does the transform of the wave function. In two dimensions, thus roughly $4N$ Fourier coefficients are needed for the description of the image transform. Secondly, each of these $4N$ coefficients enters the cross-covariance twice, since it is necessary to update

the numerator $\vec{e}\vec{s}^{*}$, in Equation (7), as well as the contrast normalization $|\vec{s}|$ in the denominator, whereas the normalization $|\vec{e}|$ of the experimental image is fixed.

In summary, the computation time required for the calculation of the evaluation function from scratch is roughly proportional to $(2N^2 + 8N)$, whereas the computation time required for the updating technique is roughly proportional to $12N$. The gain in computational speed can be roughly estimated by the corresponding ratio γ, with

$$\gamma \approx (N + 4) / 6 \qquad (9)$$

An 8-beam reconstruction is thus accelerated approximately only by a factor of two, whereas a 56-beam reconstruction can be accelerated already by an order of magnitude when employing the updating technique instead of a calculation from scratch.

Representation of the EPW

The Fourier-space representation of the EPW (Eqn. 1) can be denoted formally as a sequence of amplitude and phase values. Given a periodic wave function which is defined by a discrete and finite set of N Fourier coefficients (or beams) $F(\vec{k})$, the wave function can be encoded as a sequence w, consisting of $2N$ components given by

$$w = \{ a_1, a_2, ..., a_N, \phi_1, \phi_2, ..., \phi_N \} \qquad (10)$$

where the a_i denote the amplitude of the ith coefficient and the ϕ_i the phase, respectively. The special arrangement of amplitude and phase values within the sequence w is purely arbitrary, a different choice would be just as good. It is only important to know that a particular position represents an amplitude or a phase value belonging to a particular wave vector. Moreover, it is not necessary to use an amplitude-phase notation based on decimal numbers. Any parameterization of the wave function is possible as long as completeness and uniqueness is guaranteed. For example, the wave function could be displayed alternatively as a stream of binary digits.

The goal of the stochastic search is to modify an initial sequence w_i filled with random numbers in such a way that the images calculated from an output sequence w_o finally match exactly the experimental images. In the context of Simulated Annealing, the sequence w to be optimized is called a configuration, whereas in the terminology of genetic algorithms, a particular sequence w represents a chromosomal sequence of genes describing the properties of an individual.

Simulated Annealing

The technique of Simulated Annealing is derived from fundamental principles of statistical mechanics. Basic components of this technique were introduced by Metropolis and coworkers (Metropolis *et al*., 1953) who considered a system of particles in thermal equilibrium. The further development of this technique was achieved by Kirkpatrick (Kirkpatrick *et al*., 1983; Kirkpatrick, 1984), who applied it successfully to problems of combinatorial optimization.

The principle of Simulated Annealing is based on the observation that a many-particle system, which is in thermal equilibrium at a given temperature T, adopts a minimum of its free energy. The search for a global minimum of the evaluation function $E(\Psi)$ (Eqn. 8) belonging to a N-beam problem can be treated analogously to the search for a global minimum of the free energy belonging to a system of N interacting particles.

In statistical mechanics, the probability π_s of finding a system in a configuration s is given by the Boltzmann distribution

$$\pi_s = \frac{e^{(-\frac{E_s}{kT})}}{\sum_{w \varepsilon S} e^{(-\frac{E_w}{kT})}} \qquad (11)$$

the system is in thermal equilibrium. In the above equation, S denotes the set of all possible configurations, E denotes the energy belonging to a certain configuration, T denotes the temperature and k is Boltzmann's constant. A transition from an initial configuration i to a final configuration f occurs with a probability $p_{i \rightarrow f}$ where

$$p_{i \rightarrow f} = \frac{\pi_f}{\pi_i} = e^{-\frac{(E_f - E_i)}{kT}} \qquad (12)$$

A transition always occurs for $E_f \leq E_i$, otherwise the transition occurs with a probability $p < 1$. The thermodynamic system is located in configuration space on a shell which has a width proportional to kT. A decrease of the temperature leads to a decrease of the surface as well as of the width of this shell. In the limit $T \rightarrow 0$, the system arrives at a set of configurations, or even a single configuration, which has the lowest possible energy. Such a configuration, if unique, is called the ground state.

The Simulated Annealing algorithm consists of the following components: a representation of the configuration to be optimized (Eqn. 10), a generator for an initial configuration, a generator for new configurations, an evaluation function E (Eqn. 8), which rates a given configuration, a function calculating the transition probability $p_{i \rightarrow f}$ (Eqn. 12) and a temperature schedule describing the decrease from an initial temperature to a final temperature close to zero.

Concerning the present purpose of the reconstruction of the exit-plane wave function, an initial configuration of amplitude and phase values (Eqn. 10) is set up by means of a random number generator. A new trial wave function is established by replacing either the amplitude or the phase

value of a single randomly chosen beam by a random number. The "energy" values of the initial and the altered wave functions are calculated by means of the dimensionless evaluation function given in Equation (8). The transition probability from the initial wave function to the altered wave function is calculated by means of Equation (12), where the temperature factor $1/kT$ is substituted by a dimensionless quantity. The transition is accepted if the probability is larger than a random number generated between 0 and 1, otherwise the transition is rejected. After a certain number of successful transitions, the current configuration is assumed to represent an equilibrium state, and the temperature T is lowered following a law $T_{n+1} = cT_n$, with $c < 1$. This procedure is repeated until the temperature has fallen close to zero. The logical sequence of the Simulated Annealing algorithm is shown in a compact form in Figure 1.

During one program run, the search behavior of the algorithm is controlled by the temperature schedule and varies continuously between random search and local search. For very high temperatures, if kT is much larger than the possible bandwidth of $|E_f - E_i|$ of 180°, almost any transition from an initial to a final state is accepted (Eqn. 12) and good estimates of the wave function found so far are "forgotten" and replaced by worse configurations and vice versa. On the other hand, in the case where $kT = 0$, only "energies" E_f are accepted which are smaller than the previous value E_i. By keeping $kT = 0$ fixed during the complete program run, the Simulated Annealing algorithm can thus be used alternatively as a conventional stochastic hill climbing algorithm.

The Simulated Annealing algorithm is not well suited for fine-tuning amplitude and phase values which are already very close to the optimum solution. Global search is continued until the very end of the optimization process, and a large part of the computational effort is wasted. One possibility for avoiding this waste of computation time is the parallel implementation of local search characteristics without changing to a different type of algorithm. Convergence is accelerated by the application of additional configuration changes, which are centered around the nearly optimum amplitude and phase values found so far. Since these additional configuration changes are executed in parallel to the regular changes, the generality of the algorithm is not violated.

Genetic algorithms

The principles of genetic algorithms are derived from a simplified model of evolution, which can be considered as a search process occurring in animated nature. The goal of this search process is the optimization of a species with respect to a given environment. Genetic algorithms were first developed by Holland (1975), who found that under certain conditions genetic algorithms converge to nearly

```
set initial value for temperature;
choose initial configuration;
calculate evaluation function for initial configuration;
repeat
   repeat
     generate new configuration;
     calculate evaluation function for new configuration;
     calculate transition probability for new configuration;
     decide whether new configuration is accepted;
     if accepted:
       replace current configuration by new one;
   until thermal equilibrium is reached;
   decrease temperature by a small amount;
until ground state is reached.
```

Figure 1. Compact description of the Simulated Annealing algorithm.

```
create an initial population of individuals;
rate members of the population by an evaluation function;
repeat
   with a probability p(cross):
     choose two members of the population;
     carry out a crossing over operation with these members;
     rate the descendant;
     replace a member of the population by the descendant;
   with a probability p(mutation):
     choose a member of the population;
     choose a gene in the candidate's chromosome;
     change it randomly;
     rate the candidate;
until species is optimized.
```

Figure 2. Compact description of genetic algorithms.

optimal solutions.

A simple model of evolution, which is sufficient for the solution of optimization problems, consists of the following components: a chromosomal representation of the properties to be optimized (Eqn. 10), a set of individuals forming a species where each individual represents a different chromosomal configuration, an evaluation function which measures the fitness of individuals with respect to the environment (Eqn. 8) and genetic operators, such as mutation and crossing over.

In genetic algorithms, the mutation operator changes a single gene or a small group of genes in a certain chromosome. In this way, new properties are introduced into the population. Crossing over is often carried out by copying a sequence of genes from one chromosome to

another, thus combining existing information in a new way. The probability that a member of the population is selected for crossing over and for creating an offspring is chosen proportional to its fitness. The crossing over operation creates the power of genetic algorithms, since it is likely that the combination of two fairly good solutions can form a better one. The search is carried out in an implicitly parallel fashion by exploiting the wealth of information stored in the entire population. High performance genes become building blocks for new chromosomes and are propagated as a knowledge base from generation to generation.

In our genetic algorithm, the chromosomal representation of the exit-plane wave function is a sequence of amplitude and phase values defining the Fourier coefficients of the wave function (Eqn. 10). An initial population is created by filling the chromosomes of all individuals with uniformly distributed random numbers. Mutation is carried out in a similar way to the configuration change performed in the Simulated Annealing algorithm. Uniformly distributed random numbers are used for the choice of a certain individual, for the choice of a gene within its chromosomal representation and for the replacement of the stored amplitude or phase value by a new one. The crossing over operation is also guided by random events. The first parent for crossing over is always the best solution present in the population; the second parent is selected randomly among the remaining members of the population. The sequence of amplitude and phase values which is transferred between the two parents forming a new individual is allowed to vary in its length from just one gene up to a complete chromosomal sequence. The length as well as the end points of this sequence are determined by random numbers. In order to keep the population size constant, an offspring generated by the crossing over operator replaces the currently "worst" individual of the population. Since there are a large variety of possibilities for establishing genetic systems, the compact description shown in Figure 2 is given in a more general form.

In contrast to the Simulated Annealing algorithm, the search properties of our implementation of the genetic algorithm do not change during one program run. The search behavior is tuned via a fixed mutation rate. In the case that the individuals' genes are changed more often by mutation than by crossing over, random search characteristics dominate. Such behavior is similar to that of Simulated Annealing during its initial high-temperature phase. On the other hand, a low mutation rate leads to a local optimization of the already found solutions at the cost of global search. This situation corresponds to Simulated Annealing at low temperatures. The convergence behavior depends also on the number of individuals involved in the genetic optimization. Since the mutation rate and the population size are fixed, the success of the genetic algorithm depends much more sensitively on the choice of these fixed parameters than the Simulated Annealing algorithm depends on its continuous temperature schedule.

Similar to the Simulated Annealing algorithm, the genetic algorithm is very inefficient in fine-tuning solutions which are close to the optimum. In order to improve the speed of convergence, a strategy was developed which is analogous to the strategy chosen for the Simulated Annealing algorithm. Convergence is accelerated by applying extra mutations which are centered around the already nearly optimum amplitude and phase values.

Apart from the difference in search properties, there is a further difference between the Simulated Annealing algorithm and the genetic algorithm applied here. In the present context, the Simulated Annealing technique allows implementation of the highly efficient updating technique for the calculation of the evaluation function, which has been explained earlier in this paper. This is possible in the case of Simulated Annealing, since only one amplitude or phase value of the wave function is changed at a time. In contrast, the crossing over operator of the genetic algorithm requires at the same time the exchange of a large number of amplitude and phase values between two parents. Since an offspring consists on average of 50% of the genes of parent 1 and 50% of the genes of parent 2, an updating based on either parent 1 or parent 2 is of very limited advantage.

Performance Tests

In order to check the reliability and accuracy of the described algorithms, simulated images were used as a test input. The use of simulated images has the benefit that the output of the reconstruction algorithms can be compared directly with the EPW which was used to generate the input images. In contrast, the correct EPW is unknown in experimentation and such a direct comparison is not possible.

Simulated images of the ordered intermetallic alloy Ni_4Mo have been chosen for test purposes. This alloy is based on a fcc (face-centered cubic) matrix with a lattice constant of 0.36 nm. A projection of 2 x 2 unit cells of the $D1_a$ superstructure along [001] is shown in Figure 3. The typical sequence of four Ni atoms followed by one Mo atom along a <100> fcc direction as well as the 18° rotation between the projected fcc and $D1_a$ unit cells can be seen there.

The EPWs shown in this paper which served as input for image simulations were calculated by means of the EMS package (Stadelmann, 1987). The electron optical imaging process was simulated by means of a self-written routine. This routine has been cross-checked with the corresponding EMS routine IM1 and produces identical results.

Two test images comprising 4 x 4 $D1_a$ unit cells are displayed in Figure 4. The simulations are based on an

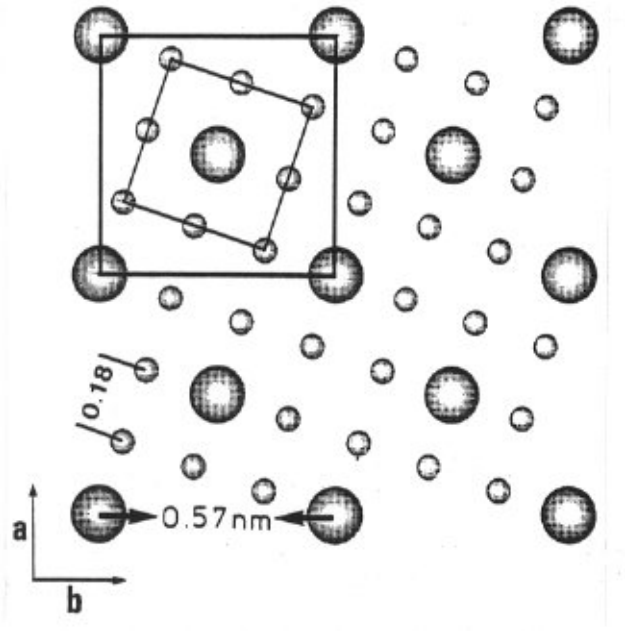

Figure 3. Projection of 2 x 2 $D1_a$ unit cells along the [001] zone axis. Mo atoms are indicated by large circles and Ni atoms by small circles. The $D1_a$ unit cell is marked by thick lines and the fcc unit cell by thin lines.

accelerating voltage of 400 kV and a spherical aberration constant of C_s of 1 mm. For the defocus spread, a value of 10 nm was used, and for the semi-convergence angle of illumination, a value of 0.9 mrad was used (1/e-values). The aperture radius used for the simulations is 7.5 nm^{-1}. The specimen thickness belonging to the two images shown in Figure 4 is 25 nm, the defocus value for image 1 is -35 nm, and that for image 2 is -70 nm. The relatively high specimen thickness has been chosen in order to obtain strongly diffracted beams which enhance the nonlinear contrast formation. From both images, it is not possible to deduce the projected atom positions by simple visual inspection.

Reconstructions of the EPW from the two images shown in Figure 4 were performed in two different modes. In the 29-beam mode, all beams within the objective aperture were treated independently from each other. In the 8-beam mode, the knowledge about the four-fold symmetry of the projected structure was exploited, and only a set of 8 (28/4 + 1 = 8) beams not related by symmetry had to be retrieved.

The input EPW used for the simulation of the two images of Figure 4 is displayed in Figure 5a, together with the output EPWs retrieved by the Simulated Annealing algorithm (Fig. 5b) and by the genetic algorithm (Fig. 5c). The typical four-by-one sequence of Ni and Mo atoms, which is not visible in the input images of Figure 4, is clearly revealed by the phase minima of the displayed EPWs. The

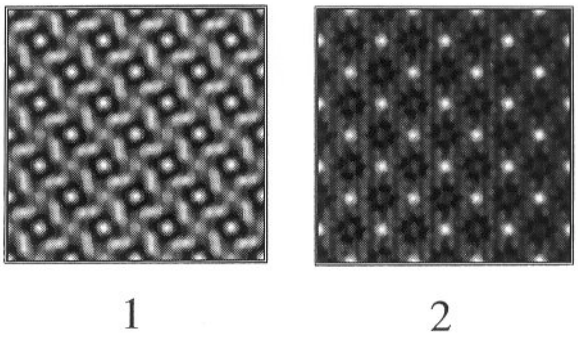

Figure 4. Simulated high-resolution images of Ni_4Mo viewed along the [001] zone axis. Exactly 4 x 4 unit cells are shown. Image 1 is simulated for a specimen thickness of 25 nm and a defocus value of -35 nm; image 2 is simulated for the same specimen thickness and a defocus value of -70 nm. The corners of the images coincide with Mo positions.

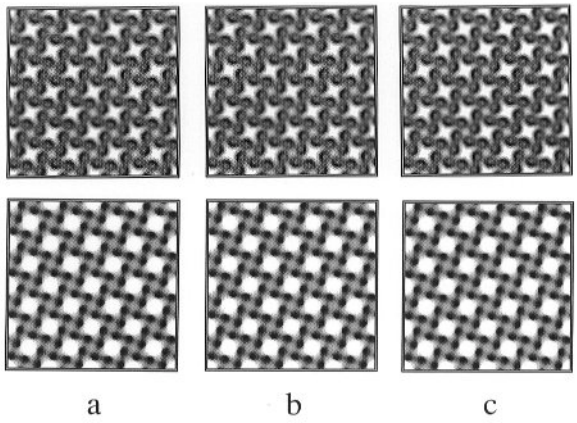

Figure 5. EPW of [001]-oriented Ni_4Mo at a specimen thickness of 25 nm. Each image comprises exactly 4 x 4 unit cells. The images in the top row show the amplitude, those in the bottom row show the phase of the wave function. (**a**) Generating EPW which was used to simulate the model images of Figure 4. (**b**) EPW reconstructed from the model images by means of the Simulated Annealing algorithm in the 29 beam mode. (**c**) EPW reconstructed by means of the genetic algorithm in the 29 beam mode.

reconstructions belonging to the wave functions of Figures 5b and 5c were both performed in the 29-beam mode. The reconstruction accuracy is so high that no difference between input and output wave functions can be observed visually in Figure 5. In order to be able to resolve small

differences between input and output wave functions, the reconstruction results are compared numerically with the correct solution in Table 1 for the Simulated Annealing algorithm and in Table 2 for the genetic algorithm. For the sake of compactness, only non-symmetry-related beams resulting from reconstructions in the 8-beam mode are displayed there. Except for the {400} beams and the {330} beams, the relative errors in the amplitudes are typically less than 0.1%, and the absolute errors in the phases are typically less than 0.02°. These extremely small errors are not caused by the algorithms themselves but by the numerical accuracy limit of the programs in use. Whereas typical values of the evaluation function (Eqn. 8) are around 90 degrees at the beginning of a program run, convergence terminates at values near 10^{-2} degrees, which is close to the minimum image difference that can be displayed as a cross-covariance using single precision numbers. Since the convergence is finally terminated by the number precision and not by the algorithms themselves, even smaller pattern differences between simulated and experimental images could be exploited by the use of double-precision numbers. In practice, image noise and other experimental inconsistencies overwhelm such small pattern differences already by orders of magnitude and an increase of the number precision in use would be completely overdone.

The reason that the {330} and {400} beams are retrieved with comparatively larger errors can be explained by the fact that the related spatial frequencies lie close to the information limit of the microscope. Due to the dampening effects caused by the partially coherent illumination, the contribution of these beams to the image contrast can be hardly detected even in ideal noise-free simulations. A more detailed investigation on the achievable accuracy under partially coherent illumination conditions, a discussion of the convergence properties and of the uniqueness of the results can be found in the work by Thust *et al*. (1994).

In order to compare qualitatively the search efficiency of both algorithms, the fast updating technique tailor-made for the Simulated Annealing algorithm was not employed at first. Using identical subroutines for image calculation and comparison, both algorithms need approximately the same CPU (central processing unit) time for the solution of the phase retrieval problem discussed here. On a DEC 3500 ALPHA (Digital Equipment Corp., Maynard, MA) workstation, the solution of the 8-beam case requires between 1 and 2 minutes of CPU time, the 29-beam case takes approximately 10 minutes, independently of the type of algorithm in use. This result indicates that the search efficiency of both algorithms is roughly equal. Owing to the possibility of using the fast updating technique, the 29-beam problem can be solved with the Simulated Annealing algorithm within 2 minutes of CPU time, which is in good agreement with the estimate given by Equation (9).

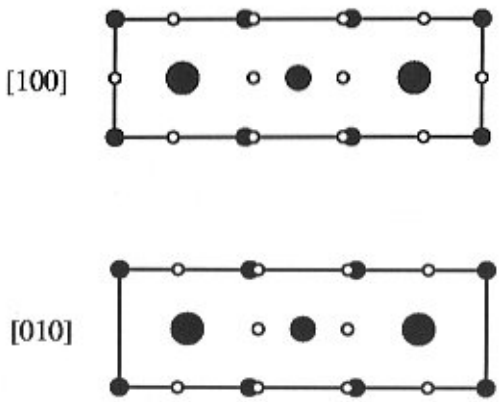

Figure 6. Top: Projection of a $YBa_2Cu_3O_7$ unit cell along the [100] zone axis. Bottom: Projection along the [010] zone axis. Full, large circles denote Ba, medium denote Y and small denote Cu positions. Oxygen positions are marked by open circles. The projected unit cell has a size of 1.17 x 0.39 nm^2 ([100] projection).

Application example

In the previous section, it was shown that phase retrieval problems up to 29 Fourier coefficients can be solved readily by means of stochastic algorithms, even though only two through-focus images were used as input. It is interesting to investigate the question if two-image reconstructions involving an even higher number of beams can still be tackled by the discussed algorithms. To answer this question, the high-T_c superconductor $YBa_2Cu_3O_7$ was chosen as a test object. Due to the large unit cell occupied by four different atom species, the number of Fourier coefficients needed for the synthesis of the wave function is relatively high. Projections of the unit cell along the [100] and [010] directions, which are difficult to distinguish in experiment, are shown in Figure 6. Since this distinction is not of importance in the following, the zone axis will be addressed as the [100] axis.

Experimental images were recorded with a JEOL 4000EX (JEOL, Tokyo, Japan) electron microscope operated at an accelerating voltage of 400 kV. An objective aperture with a reciprocal radius of 5.5 nm^{-1} was used for the observation of $YBa_2Cu_3O_7$ along the [100] zone axis, which results in 53 beams contributing to the EPW. Except for the aperture radius, all optical parameters correspond to those used in the previous section for image simulations.

Figures 7 and 8 show two experimental images of $YBa_2Cu_3O_7$ which were taken with different defocus from

Table 1. Reconstruction accuracy of the Simulated Annealing algorithm. Comparison of the EPW used to generate the input images of Figure 4 with the EPW obtained by the application of the Simulated Annealing algorithm. The amplitudes and phases belonging to the "input" EPW are denoted by the subscript "i," those belonging to the output EPW are marked with the subscript "o." The amplitude of the unscattered beam is normalized to the value 1, and the common phase factor has been chosen to result in the phase value zero for the unscattered beam.

(h k l)	A_i	A_o	$\Delta A/A_i$ [%]	φ_i [deg]	φ_o [deg]	$\varphi_o - \varphi_i$ [deg]
(0 0 0)	1.0000	1.0000	0.00	0.00	0.00	0.00
(1 1 0)	0.3341	0.3341	0.00	83.54	83.55	0.01
(2 0 0)	0.2481	0.2482	0.04	63.43	63.43	0.00
(2 2 0)	0.1287	0.1286	-0.08	30.48	30.48	0.00
(3 1 0)	0.1062	0.1062	0.00	25.38	25.38	0.00
(1 3 0)	0.2542	0.2542	0.00	223.99	224.00	0.01
(4 0 0)	0.0386	0.0384	-0.56	227.24	226.93	-0.31
(3 3 0)	0.0111	0.0103	-7.10	233.71	223.79	-9.92

Table 2. Reconstruction accuracy of the genetic algorithm. Comparison of the EPW used to generate the input images of Figure 4 with the EPW obtained by the application of the genetic algorithm. The amplitudes and phases belonging to the "input" EPW are denoted by the subscript "i," those belonging to the output EPW are marked with the subscript "o." The amplitude of the unscattered beam is normalized to the value 1, and the common phase factor has been chosen to result in the phase value zero for the unscattered beam.

(h k l)	A_i	A_o	$\Delta A/A_i$[%]	φ_i [deg]	φ_o [deg]	$\varphi_o - \varphi_i$ [deg]
(0 0 0)	1.0000	1.0000	0.00	0.00	0.00	0.00
(1 1 0)	0.3341	0.3340	-0.03	83.54	83.52	-0.02
(2 0 0)	0.2481	0.2479	-0.06	63.43	63.43	0.00
(2 2 0)	0.1287	0.1289	0.15	30.48	30.50	0.02
(3 1 0)	0.1062	0.1063	0.04	25.38	25.39	0.01
(1 3 0)	0.2542	0.2540	-0.06	223.99	223.98	-0.01
(4 0 0)	0.0386	0.0399	3.36	227.24	227.59	0.35
(3 3 0)	0.0111	0.0070	-36.77	233.71	240.32	6.61

approximately the same specimen area. The images have been recorded on photographic plate and were digitized from the negative by means of a CCD (charge coupled device) camera (for details, see Thust and Urban, 1992). The areas chosen for further processing are marked by a white frame in Figures 7 and 8 and contain exactly 3 x 9 $YBa_2Cu_3O_7$ unit cells. It is very likely that these frames are not located at exactly the same specimen area, since it is difficult to identify the same position on two different negatives. The maximum spatial separation between the two frames displayed in Figures 7 and 8 is estimated to be around 3 nm. Since no significant long range variation of the image patterns can be observed within the displayed areas, a spatial shift between the chosen frames is only of minor importance for the reconstruction in the case of a periodic object. It is, however, of importance that this shift corresponds to integer multiples of the lattice vectors defining the $YBa_2Cu_3O_7$ unit cell. In the following, the image extracted from the frame displayed in Figure 7 is called image 1; the image extracted from the frame of Figure 8 is called image 2.

Figure 9 (top) shows images 1 and 2 after spatial averaging over all unit cells within the extracted frames. Although each unit cell is identical after averaging, the display of 3 x 9 unit cells is maintained in order to facilitate visual pattern recognition. Before starting a reconstruction based on the averaged images 1 and 2, two aspects deserve special attention. Firstly, it is *a priori* not clear, which position of unit cell 1 corresponds to a particular position in unit cell 2, i.e., a common origin has to be found. This

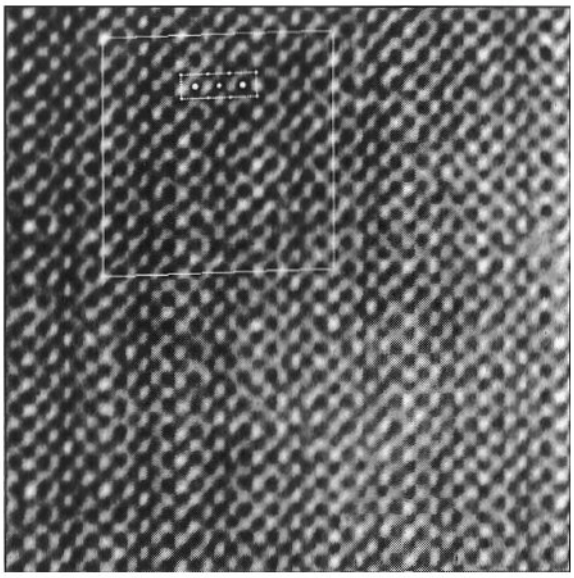

Figure 7. High-resolution image of $YBa_2Cu_3O_7$ taken along the [100] zone axis. The image was recorded with a CCD camera from the photographic negative and consists of 512 x 512 pixels. The area chosen for reconstruction is surrounded by a frame of size 3.5 x 3.5 nm^2. A projected unit cell of $YBa_2Cu_3O_7$ is marked within the frame.

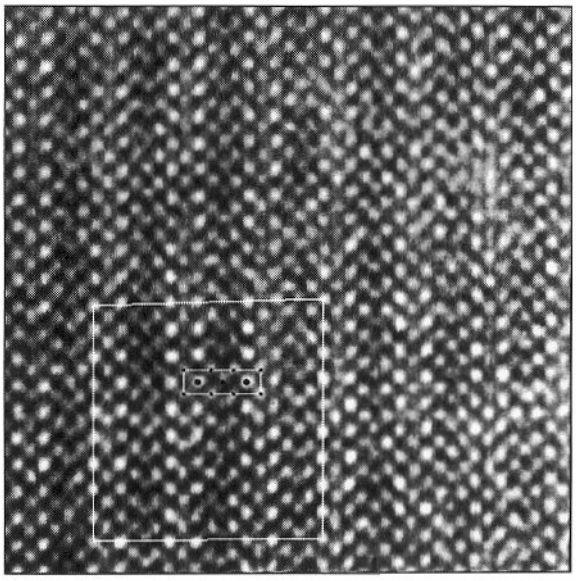

Figure 8. High-resolution image of $YBa_2Cu_3O_7$ taken from the same specimen area as the image of Figure 7, but at a different defocus setting. See legend of Figure 7 for details.

problem can be solved by means of symmetry arguments except for one ambiguity. It is possible to overlay intensity maxima of image 1 with such of image 2, or alternatively, by a half-unit-cell shift along the [010] direction (y-direction in Fig. 9) with intensity minima of image 2. A second problem arises due to the fact that neither the exact defocus values of the images 1 and 2, nor the relative focal difference between them are known with sufficient precision. In order to solve these problems, image simulations were carried out with the automatic image-scan routine described in Thust and Urban (1992). For different choices of the unit-cell origin, the best fitting simulated images with respect to defocus and specimen thickness were automatically determined.

The simulated images yielding the best fit to the experimental images 1 and 2 are displayed at the bottom of Figure 9. These images, which are called image 3 and 4 in the following, belong to a specimen thickness of 3.5 nm and to defocus values of -30 nm and -72 nm, respectively. On visual standards, the correspondence between simulation and experiment is fairly good, although not convincing. Whereas the main contrast features, such as number and position of minima and maxima are in good correspondence, the experimental images appear somewhat blurred compared to the simulations. This discrepancy is revealed prominently when using the evaluation function defined in Equation (8) as a quantitative measure. Since the experimental images 1 and 2 deviate slightly from the mm symmetry of the ideal structure, these images were symmetrized before the comparison with the perfectly symmetric simulations (Thust and Urban, 1992). After symmetry correction, a relatively large residual value of 28 degrees is obtained for the evaluation function.

In order to check the capability of stochastic algorithms to solve the specific reconstruction problem, at first not the experimental images 1 and 2, but the simulated counterparts 3 and 4 (Fig. 9) were used as input for the reconstruction. For reasons of flexibility and speed, the reconstructions presented in this section were performed exclusively with the Simulated Annealing algorithm. With 53 beams within the objective aperture, a reconstruction down to values of the evaluation function close to 10^{-2} degrees takes less than 10 minutes of CPU time on a DEC 3500 ALPHA workstation.

The reconstructed EPW is displayed in Figure 10 together with the "correct" EPW used as input for the simulation of the images 3 and 4. It can be seen by visual comparison that the correspondence between input and output EPW is perfect. A quantitative comparison of the wave functions yields an even better correspondence than that shown in Tables 1 and 2 for the Ni_4Mo calculations. The fact that the reconstruction from simulated images of $YBa_2Cu_3O_7$ is more precise, even though the number of beams is almost twice as high, can be explained by the smaller

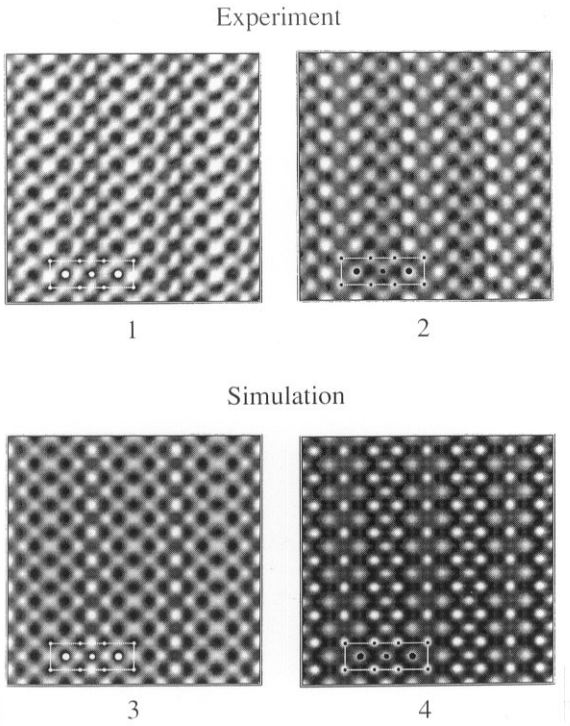

Figure 9. (**1**) Image containing 3 x 9 identical unit cells of $YBa_2Cu_3O_7$, which is obtained after averaging over the unit cells surrounded by the frame in Figure 7. (**2**) Image obtained after averaging over the unit cells surrounded by the frame in Figure 8. (**3**) Simulated image which fits best to image 1. (**4**) Simulated image which fits best to image 2. A projected $YBa_2Cu_3O_7$ unit cell is marked in the images.

aperture used for the simulation of the $YBa_2Cu_3O_7$ images. The aperture with a reciprocal radius of 5.5 nm^{-1} excludes such beams from the imaging process which lie close to the information limit of the microscope. As has been explained in the previous section, the contrast contributions of such beams are extremely small and may fall below the numerical detection limit.

It was further investigated if a premise-free determination of the relative focal distance between two images is possible by means of the Simulated Annealing algorithm. Again, simulated images were used first to test the feasibility of such a determination. Several trial reconstructions with different assumptions about the focal distance between the simulated images 3 and 4 of Figure 9 were made. It can be seen in Figure 11 that the smallest obtainable value of the evaluation function increases strongly in the case that the reconstruction is based on a wrong assumption about the focal distance. The correct value of 42 nm is revealed unambiguously by a steep minimum of the evaluation function.

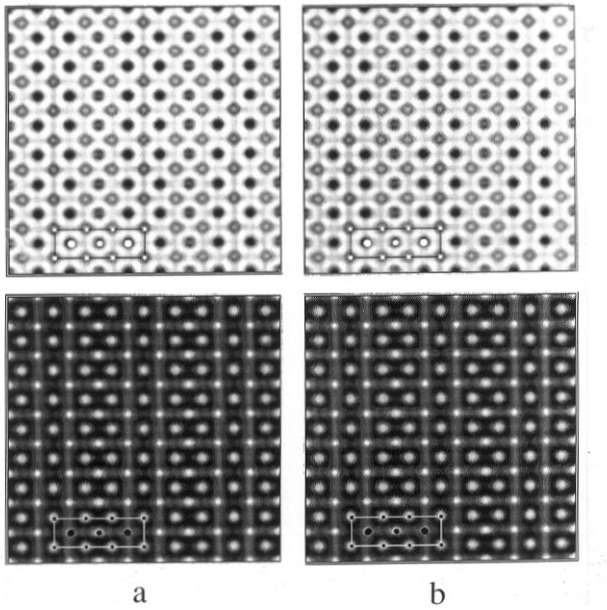

Figure 10. (**a**) Amplitude (top) and phase (bottom) of the EPW reconstructed from the simulated images 3 and 4 of Figure 9. (**b**) Amplitude (top) and phase (bottom) of the EPW which was used as input for the simulation of the images 3 and 4 of Figure 9. The resolution of the displayed EPWs is limited to 5.5 nm^{-1}. The common phase factor is chosen to result in the phase value zero for the unscattered beam. A projected $YBa_2Cu_3O_7$ unit cell is marked in all images.

After having assured that a determination of relative focal distances is possible from simulated images, the same procedure was applied to the experimental images 1 and 2 shown in Figure 9. Similar to the simulations, a single minimum of the evaluation function is found in the experimental case. The value of the focal distance between images 1 and 2 is 36 nm and differs by 6 nm from the value determined from the simulated images 3 and 4. Whereas in the simulated case the minimum of the evaluation function is very sharp and has a value of approximately 10^{-2} degrees, the experimentally determined minimum is much broader and has a value of 8 degrees (Fig. 11).

Amplitude and phase of the wave function resulting from the reconstruction of the experimental images 1 and 2 are shown in Figure 12. The projected positions of the cations Y, Ba, and Cu are clearly revealed in both the amplitude and the phase image.

As is expected for a thin, weakly scattering object, the atomic positions are revealed by minima in the amplitude and by maxima in the phase image. Whereas the asymmetry of the input images is dominant in the amplitude image, the phase image yields an almost distortion-free

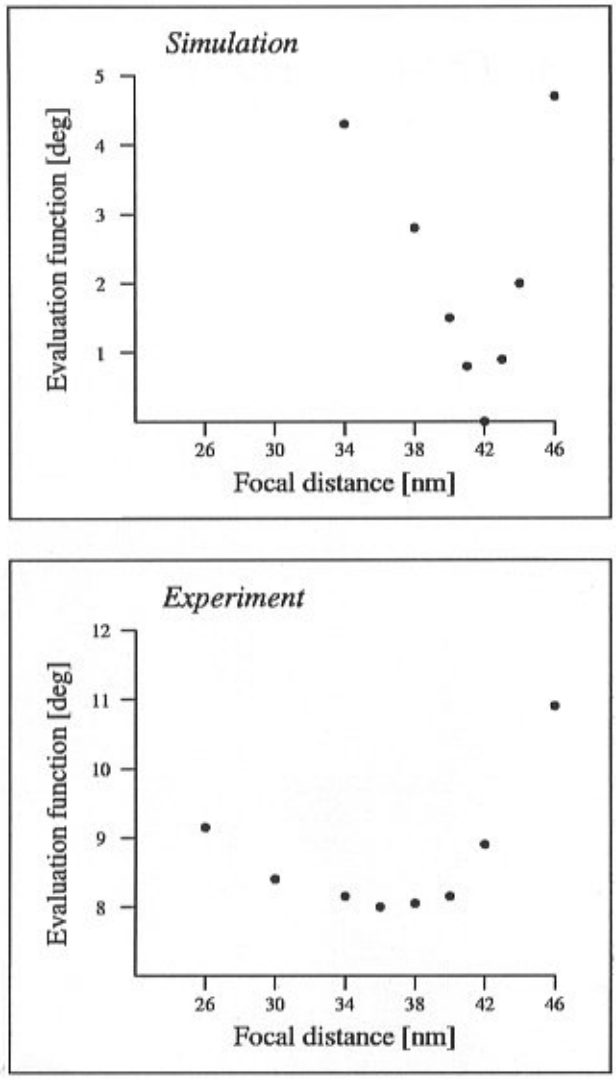

Figure 11. Display of the smallest achievable value of the evaluation function in dependence of the assumed focal distance between two images. The upper graph refers to reconstructions based on the simulated images 3 and 4 of Figure 9, the lower graph refers to reconstructions based on the experimental images 1 and 2 of Figure 9.

projection of the cation structure. A general feature of the experimentally reconstructed EPW is a lack of high-resolution detail when compared with the simulation of Figure 10. As is the case for the simulated EPW of Figure 10, the projected oxygen positions atoms are not directly visible in the experimental reconstruction.

The tests performed in parallel with the simulated images allow judgement of the reliability and precision of the reconstruction technique directly on the level of the EPW. In experimentation, such a direct check is not possible. Nevertheless, the quality of the reconstructed EPW can be judged indirectly by a comparison of the images re-calculated from the EPW with the experimental input images. Such a comparison is shown in Figure 13. (In this context, it is of importance to distinguish between "simulated" and "recalculated" images: simulated images are based on a known structure model and result from a calculation of the dynamic electron scattering process and of the subsequent electron optical imaging process. In contrast, recalculated images result purely from a simulation of the electron optical imaging process based on a reconstructed EPW.) The difference between the experimental input images and the recalculated images is hardly visible by eye in Figure 13, which means that the EPW displayed in Figure 12 reproduces almost perfectly the experimentally observed image contrast. The corresponding value of the evaluation function is 8 degrees, whereas a comparison of the experimental images with the best fitting simulated images (Fig. 9) yields a much higher value of 28 degrees. Although such an indirect comparison on the image level cannot be seen as a strict proof, the obtained results indicate that the reconstructed EPW displayed in Figure 12 is of considerably higher relevance with respect to the experiment than is the EPW displayed in Figure 10, which was obtained via the comparison with simulations.

There are various reasons which could be responsible for the observed discrepancies between reconstruction and simulations. Although a detailed treatment is out of the scope of the present paper, some of the more evident causes are discussed in short. Compared to the simulations, the most prominent discrepancy observed in experiment is the lack of high-resolution detail. This can be already observed by eye when comparing the images of Figure 9, and is revealed quantitatively in Figure 11 by the slow variation of the experimental evaluation function with the assumed focal distance. The sharp minimum of the simulated counterpart can only be synthesized in the presence of high-frequency beams which have a small focal repetition period $\Delta Z = 2/(\lambda g^2)$. Apart from instrumental instabilities, a lack of high-resolution detail in the experimental images could be due to static atom displacements and a partial amorphization of the specimen caused either by ion milling during sample preparation or by electron irradiation during observation. Such effects have not been taken into account in our simulations. Another point to discuss is the fact that individual unit cells in the experimental input images of Figures 7 and 8 differ from each other in image contrast. A spatial averaging, as applied in this work, is not justified in a strict sense, since local structural variations cannot be treated as random noise. Therefore, the spatial averaging causes a degradation of non-periodic high-resolution detail. However, possible problems arising from the violation of

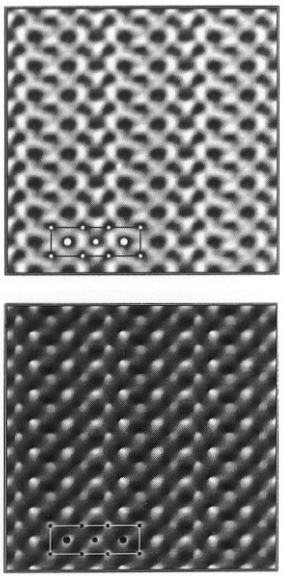

Figure 12. Amplitude (top) and phase (bottom) of the EPW obtained from the reconstruction based on the experimental images 1 and 2 of Figure 9. The common phase factor is chosen to result in the phase value zero for the unscattered beam. In both images, a projected $YBa_2Cu_3O_7$ unit cell is marked.

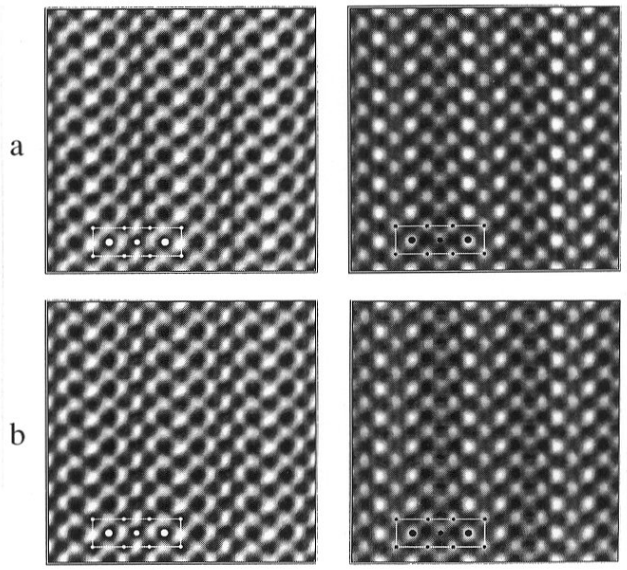

Figure 13. (**a**) Experimentally recorded and averaged images. The left image is identical with image 1 of Figure 9; the right image is identical with image 2 of Figure 9. (**b**) Images re-calculated on the basis of the reconstructed EPW displayed in Figure 12. A projected $YBa_2Cu_3O_7$ unit cell is marked in the images.

strict periodicity are not specific for the phase-retrieval procedure and arise also when comparing periodic simulations with real-life images.

Conclusions

Stochastic algorithms are highly efficient tools for the reconstruction of the exit-plane wave function belonging to periodic high-resolution electron microscopic images. Reconstructions with an accuracy close to analytical can be achieved on standard workstations within the time scale of interactive computer sessions. Whereas the classical PAM and MAL focal-series reconstruction methods seem indispensable for the solution of non-periodic problems, stochastic algorithms are superior in the case of periodic problems, because only two input images are mostly sufficient for a fully nonlinear reconstruction. Furthermore, since most high-resolution images contain periodic areas, stochastic algorithms can be applied for a premise-free determination of relative defocus intervals. Amongst the two types of algorithms applied in this work, we prefer the Simulated Annealing algorithm over the genetic algorithm. First, the Simulated Annealing algorithm can be implemented far more efficiently in this context and, secondly, the adjustment of the convergence behavior turned out to be more flexible and user friendly.

References

Coene W, Janssen G, Op de Beeck M, Van Dyck D (1992) Phase retrieval through focus variation for ultra-resolution in field-emission transmission electron microscopy. Phys Rev Lett **69**: 3743-3746.

Coene WMJ, Thust A, Op de Beeck M, Van Dyck D, Janssen AJEM, Klompstra MH (1996) Maximum-likelihood method for focus-variation image reconstruction in high resolution transmission electron microscopy. Ultramicroscopy **64**: 109-135.

Davis L, Steenstrup M (1987) Genetic algorithms and simulated annealing: An overview. In: Genetic Algorithms and Simulated Annealing. Davis L (ed). Pitman, London. pp 1-11.

Drenth AJJ, Huiser AMJ, Ferwerda HA (1975) The problem of phase retrieval in light and electron microscopy of strong objects. Optica Acta **22**: 615-628.

Frank J (1973) The envelope of electron microscopic transfer functions for partially coherent illumination. Optik **38**: 519-536.

Holland JH (1975) Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.

Ishizuka K (1980) Contrast transfer of crystal images

in TEM. Ultramicroscopy **5**: 55-65.

Kirkland EJ (1984) Improved high resolution image processing of bright field electron micrographs. I. Theory. Ultramicroscopy **15**: 151-172.

Kirkland EJ, Siegel BM, Uyeda N, Fujiyoshi Y (1985) Improved high resolution image processing of bright field electron micrographs. II. Experiment. Ultramicroscopy **17**: 87-104.

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science **220**: 671-680.
Kirkpatrick S (1984) Optimization by simulated annealing: Quantitative studies. J Stat Phys **34**: 975-986.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys **21**: 1087-1093.

Misell DL (1973) An examination of an iterative method for the solution of the phase problem in optics and electron optics: I. Test calculations. J Phys D: Appl Phys **6**: 2200-2216.

Saxton WO (1978) Computer Techniques for Image Processing in Electron Microscopy. Academic Press, New York. Section 9.7.

Saxton WO (1980) Correction of artifacts in linear and non-linear high resolution electron micrographs. J Microsc Spectrosc Electron **5**: 661-670.

Saxton WO (1986) Focal series restoration in HREM. In: Proc 11th Int Congr on Electron Microscopy (Kyoto). Post-deadline paper 1. Imura T, Maruse S, Suzuki T (eds). Jap Soc Electron Microsc, Tokyo.

Saxton WO (1993) Linear image restoration: The approaches available for real and complex objects. In: Image Interpretation and Image Processing in Electron Microscopy, Proc. Autumn School 1992 of Max Planck Institut für Mikrostrukturphysik, Halle/Saale. Heydenreich J, Neumann W (eds.). Max Planck Institut für Mikrostrukturphysik, Halle/Saale. pp. 118-128.

Saxton WO (1994) What is the focus variation method? Is it new? Is it direct? Ultramicroscopy **55**: 171-181.

Schiske P (1973) Image processing using additional statistical information about the object. In: Image Processing and Computer-aided Design in Electron Optics. Hawkes PW (ed). Academic Press, London. pp 82-90.

Stadelmann PA (1987) EMS - A software package for electron diffraction analysis and HREM image simulation in materials science. Ultramicroscopy **21**: 131-146.

Thust A, Urban K (1992) Quantitative high-speed matching of high-resolution electron microscopy images. Ultramicroscopy **45**: 23-42.

Thust A, Lentzen M, Urban K (1994) Non-linear reconstruction of the exit plane wave function from periodic high-resolution electron microscopy images. Ultramicroscopy **53**: 101-120.

Thust A, Coene WMJ, Op de Beeck M, Van Dyck D (1996a) Focal-series reconstruction in HRTEM: Simulation studies on non-periodic objects. Ultramicroscopy **64**: 211-230.

Thust A, Overwijk MHF, Coene WMJ, Lentzen M (1996b) Numerical correction of lens aberrations in phase-retrieval HRTEM. Ultramicroscopy **64**: 249-264.

Van Dyck D, Op de Beeck M (1990) New direct methods for phase and structure retrieval in HREM. In: Proc 12th Int Congress Electron Microsc (Seattle). Peachey LD, Williams DB (eds). San Francisco Press, CA. pp 26-27.

Van Dyck D, Op de Beeck M (1993) A new approach to object wave function reconstruction in electron microscopy. Optik **93**: 103-107.

Discussion with Reviewers

G. Möbus: If two images (focal values) are now considered to be sufficient to retrieve the exit wave function, why is one image still not enough?

Authors: Although we have no general proof, we found that a single input image can, at least in special cases, be sufficient to retrieve the EPW uniquely. By using only image 2 shown in Figure 4, it is possible to retrieve the corresponding EPW by means of the Simulated Annealing algorithm in the 8-beam mode. The uniqueness and the accuracy of the solution were checked with extreme care in this special case: twenty different program runs using different random numbers yielded identical results, indicating that the retrieved EPW is unique. The accuracy of the retrieved Fourier coefficients was surprisingly of the same quality than that of the corresponding two-image reconstruction. With respect to a potential practical application, one important remark has to be made: in the case of a reconstruction from one single image, the search space is covered densely with local minima of the evaluation function, which can extend even below a value of 1°. In order to determine the EPW purely on the basis of image 2 shown in Figure 4, an extremely slow cooling rate has to be chosen, and the required numerical effort increases roughly by an order of magnitude compared to the two-image reconstruction. Under experimental conditions, such small differences of the evaluation function, which lead to the correct solution, are likely to be buried in noise, and a single-image reconstruction would most probably yield unreliable results.

G. Möbus: The comparison of performance for the MAL-method (Coene *et al*., 1996) and the new stochastic method is quite complicated, since four features have been changed at once: (i) reciprocal space instead of real space, (ii) global optimization instead of local optimization, (iii) correlation instead of chi-square of Maximum Likelihood theory, (iv) perfect crystals instead of crystal defects. What do the

authors think about mixed techniques such as a stochastic global optimization of a dislocation EPW in real space? By how many orders of magnitude would the calculation time increase (roughly)?

Authors: The number of parameters describing the EPW increases considerably when changing from Fourier to real space. Using a sampling rate of 40 pixels/ nm, the real-space description of a relatively small motif of size 1 nm x 1 nm requires already 1600 pixels. Due to the non-linear character of the imaging process, the representation of the EPW requires fortunately only half of the sampling rate used for the digitization of the input images. As a consequence, one quarter of the above calculated pixel number is actually necessary to represent the corresponding two-dimensional EPW, i.e., 400 complex numbers have to be determined in order to reconstruct only a small 1 nm x 1 nm patch.

It is important to note that this statement holds only for periodic motifs. The number of involved real-space pixels becomes even higher in the non-periodic case, since the margins of the field of view cannot be reconstructed within a distance corresponding to the point spread of the microscope (see, e.g., Thust *et al*., 1996a). Whereas the sizes of the input images and the output EPW are identical for a periodic object, the reconstructible area of a non-periodic EPW is smaller compared to the input image size. As a consequence, the numerical efficiency (computation time per image area) can be substantially lower for the non-periodic case, depending on the particular imaging conditions in use and on the total area to be reconstructed.

From the above considerations, it follows that the real-space reconstruction of non-periodic objects, which extend typically over several nanometers in size, would involve by several orders of magnitude more parameters than have been treated in the body of the paper. It is difficult to estimate the dependence of the required calculation time on the number of involved parameters, since the computational effort may increase from linear (best case, linear imaging conditions) to exponential (worst case, strongly nonlinear imaging conditions). In our experience, stochastic algorithms are not well suited for the solution of such high dimensional problems, as is the real-space treatment of a crystal defect. One could even doubt if a real-space reconstruction of a defect is still feasible using stochastic algorithms. As long as approximately twenty high-quality through-focus images are available, the choice of traditional deterministic reconstruction algorithms will be the better strategy. The real strength of stochastic algorithms lies in their capability to solve the phase problem for a small set of unknowns when only very few input data are available. This property makes stochastic algorithms superior to the alternative techniques in the field of periodic objects.

P.A. Stadelmann: What could be expected from either the genetic or stochastic algorithms when the information beyond the Scherzer resolution limit is degraded by 3-fold astigmatism? Should we not use such a technique when 3-fold correctors will be installed in HREM microscopes? For example, in the JEOL 4000 EX, 3-fold astigmatism introduces phase shifts larger than $\pi/4$ for spatial frequencies close to its Scherzer resolution limit.

M. Hÿtch: The exit wavefunction reconstruction for the experimental case is convincing but the experimental conditions are assumed to be known. What happens in the presence of misalignments like beam tilt or astigmatism? Would it not be possible to include these as extra parameters?

Authors: Of course, the neglect of anisotropic aberrations like beam tilt or 3-fold astigmatism, which is present on most current ultra-high resolution instruments, will affect the reconstructed EPW, regardless of the particular reconstruction technique in use. Two viable strategies can be employed to tackle the problem: first, an aberration-clean microscope alignment can be achieved by means of correction elements which compensate *a priori* all anisotropic aberrations like coma, 2-fold and 3-fold astigmatism, or, secondly, such aberrations can be removed a posteriori from the reconstructed EPW by the application of a numerical phase plate. Unfortunately, these techniques were not yet established during the progress of the work shown here. Concerning the hardware approach, a 3-fold astigmatism corrector is meanwhile available from Philips (Philips Electron Optics, Eindhoven, The Netherlands) for their high-resolution microscopes which allows one to avoid the occurrence of any anisotropic aberration during the experiment (Overwijk *et al*., 1997). Alternatively, the software correction approach is applied routinely in the field of phase retrieval HRTEM (Thust *et al*., 1996b).

P.A. Stadelmann: Can you comment on the importance of crystal misalignment or tilt for the success of the phase retrieval?

Authors: In the case of beam tilt or other anisotropic aberrations, the EPW is modified by the imaging system on its way from the object plane to the image plane. In the case of crystal tilt, the situation is quite different. The crystal tilt is not an instrumental artifact but is a real effect which is physically present in the object plane. It is independent from the imaging or reconstruction process and behaves like an object property. If no assumptions about the object are made, an EPW affected by crystal tilt will thus be retrieved with the same success as any other "non-tilted" EPW. Whereas the reconstruction is straightforward, the direct interpretation of "tilted" EPWs can be difficult. In the case of strong crystal tilt, it might be necessary to compare the experimentally retrieved EPW with

corresponding simulated EPWs in order to draw conclusions about the specimen structure.

P.A. Stadelmann: Did you ever try to apply these reconstruction schemes to more complicated crystal structures like the large unit cell oxide $Nb_{10}Ti_2O_{29}$, where a couple of hundreds of beams contribute to the image?
Authors: The examples of the relatively complicated unit cell of $YBa_2Cu_3O_7$ involving 53 beams are the maximum we tried out. Since we did not encounter any problems, we think that the capacity limit of the discussed algorithms has not yet been reached and that even more complicated unit cells can be tackled. It is, however, difficult to predict whether several hundred beams are still in reach for the discussed algorithms, since the feasibility would also depend on factors like the number of input images, the available computing power and the total computation time, which is still regarded as acceptable.

P.A. Stadelmann: What would be the necessary computer power in MIPS (or Mflops) to perform real-time phase retrieval using the "Stochastic Algorithm" for a typical perovskite crystal $BaTiO_3$?
Authors: It is difficult to give an exact number, since the outcome will depend on factors like the desired resolution (number of beams), the number of input images, the degree of nonlinearity involved in the imaging process and the desired accuracy of the results. Especially the last factor is of considerable importance, since the treatment of the phase retrieval problem down to a cross-covariance level of 10^{-2} degrees, which has been impressively demonstrated with simulated input images, costs much computational effort, but is completely overdone in experiment due to the existence of noise and other data inconsistencies. Reconstructions based on experimental data can thus be executed much faster, since the best obtainable values of the evaluation function are rarely smaller than approximately one degree. On a reasonably up-to-date PC or workstation ($\approx$ 30-100 Mflops), the experimental reconstruction of a simple unit cell like that of $BaTiO_3$ should require CPU times which range from a couple of seconds in the best case, to some tens of seconds in the worst case.

P.A. Stadelmann: What do you think of combining your algorithms with CRISP (crystallographic image processing) or equivalent kinematical programs?
Authors: Typical crystallographic image processing programs rely on the fact that the sample is a weakly scattering object or even a weak phase object, implying that the scattered intensity is small compared to the intensity of the incident beam. In a strict sense, it is doubtful to apply such programs to inorganic crystals important to materials science, because nonlinear imaging theory, which is important for specimen thickness values exceeding a few nanometers, is completely ignored. Moreover, the obtained results exhibit the Friedel symmetry of the input images, making an *a posteriori* software correction of optical aberrations impossible. The correct treatment of aberrations on the level of a reconstructed EPW would be clearly superior to the usual symmetry averaging procedures used in crystallographic programs, since the symmetry averaging approach is based on purely phenomenological considerations and is not correct from a mathematical point of view. We think that the incorporation of the stochastic approach into crystallographic image processing packages would be a real benefit in many aspects.

M. Hÿtch: For the solution of the experimental case, how do the contrast levels compare? It seems that again the best fitting wavefunction has the appearance of a wavefunction emerging from a very thin crystal.
Authors: Taking the zero-beam intensity as a reference, we found that the total scattered intensity is in general smaller for experimentally reconstructed EPWs than for simulated EPWs. This behaviour corresponds to findings made in image simulation, where it is was observed that the experimental image contrast can be considerably lower than the corresponding simulated contrast, taking the image mean value as a reference. Since both methods - comparison between experimental and simulated images, on the one hand and comparison of reconstructed and simulated EPWs, on the other hand - are based on the same kind of experimental input images, it is not surprising that a similar "low contrast" behaviour is observed in both cases. Such low contrast output can be obtained in simulation only for relatively small values of the specimen thickness. In our case, a wrong measurement of the image mean value due to a possible "fog" effect cannot be responsible for the discussed discrepancy, since the image mean value does not enter our reconstruction procedure (we use the cross-covariance instead of the cross-correlation). At the time being, we have no stringent explanation for this effect.

P.W. Hawkes: The absolute values of the intensities measured on experimental images are not necessarily correct, for reasons that are well understood but not easy to remedy. How will such measurement errors affect the proposed algorithms?
Authors: Indeed, the optical density of the photographic emulsion need not be related linearly to the charge density of the incident electrons. However, a nearly linear dependence can be obtained by making exposures with sufficiently low electron doses. If one nevertheless uses input images which exhibit saturation or even cut-off effects, those nonlinearities will also affect the reconstructed EPW. Interestingly, the introduced errors become smaller with an

increasing number of input images. This is due to the fact that the contrast distortions affecting the various input images constitute mutually inconsistent information which is damped out due the intrinsic averaging process inherent to the reconstruction.

Additional Reference

Overwijk MHF, Bleeker AJ, Thust A (1997) Correction of three-fold astigmatism for ultra-high-resolution TEM. Ultramicroscopy **67**: 163-170.