Image Processing Approaches to Biological Three-Dimensional Electron Microscopy

Gabor T. Herman,1 Roberto Marabini,1 José-Maria Carazo,2 Edgar Garduño,1 Robert M. Lewitt,1 Samuel Matej1

1 Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104
2 Centro Nacional de Biotecnología, Campus Universidad Autónoma, 28049 Madrid, Spain

ABSTRACT: Electron microscopy is a powerful technique for imaging complex biological macromolecules in order to further the understanding of their functions. When combined with sufficiently careful sample preparation procedures that preserve the native structure of the macromolecules and with sophisticated image processing procedures, electron microscopy can lead to very informative estimates of the three-dimensional (3D) structures of the specimens under study. 3D reconstruction from electron microscopic data is achieving high goals and exceeding expectations unthinkable only a few years ago. However, there are still some areas where either not enough work has been invested or the work has not as yet been fruitful. We describe image processing approaches that shed further light on some of these difficult areas. © 2000 John Wiley & Sons, Inc. Int J Imaging Syst Technol, 11, 12–29, 2000

Key words: electron microscopy; image processing; three-dimensional reconstruction; reconstruction from projections

I. INTRODUCTION

In this paper, we concentrate on three areas in which multidisciplinary approaches have the potential for improving three-dimensional (3D) electron microscopy (EM):

A. Incorporation of realistic image formation models into new reconstruction methods. There is a need for reconstruction algorithms that take into account realistic image-blurring models of the aberrations of the electron microscope and that are at the same time noise resistant and flexible with respect to the different data collection geometries.

B. Incorporation of knowledge regarding the specimen obtained by means other than EM; specifically, incorporation into the 3D reconstruction process of high-resolution relief information obtained by either atomic force microscopy (AFM) or high-resolution metal shadowing and incorporation of information regarding the chemical nature of the specimen.

C. Improvement of the rendering and the analysis of the reconstructed volumes by the development of more accurate segmentation (differentiation between the specimen and its background) and visualization algorithms.

These three basic areas for improving EM need to be complemented by a rigorous approach to validating claims of superiority of any of the newly developed methods over those used in current practice. The approach should include very realistic (quantum mechanical) simulations of the electron microscopic imaging process on structures deposited in the protein data bank (PDB), as well as additional work on specimens whose structures at atomic resolution is known due to alternative approaches.

II. MOTIVATION

The basic rationale for seeking structural information in biology is that the understanding of the precise way in which key reactions in a cell are carried out ultimately requires knowing the spatial interactions among the partners of that reaction, and that for such knowledge information on the 3D structure of these partners is essential. There are several complementary approaches to unraveling the 3D structure of biological macromolecules; examples are X-ray diffraction, nuclear magnetic resonance (NMR) spectroscopy, and 3D EM (and strong emerging methodologies, such as AFM).

This paper concentrates on 3D EM, which has the unique capability to work with quite large specimens in a broad range of situations, such as specimens arranged on a 2D crystal, on a helix, on an icosahedron, or just as single particles. The structural information obtained in this way seldom reaches atomic resolution, but it has a great impact in the resolution region between about 6 and 20 Å. It is, therefore, clear that 3D EM provides a perfect framework into which atomic-level structural information on parts of a macromolecular complex, coming from either X-ray diffraction or NMR spectroscopy, can be assembled.

In the following subsections, we discuss the state-of-the-art in the three areas mentioned in the Introduction, pointing out where there is a need for methodological improvements and the potential biological significance of such improvements.

A. Incorporation of Image Formation Models. The specific significance of this topic arises from the fact that the 3D reconstruction process cannot be made truly quantitative unless instrumental inaccuracies are carefully modeled and taken into consideration.
Although some work has been devoted to the field of 3D EM to this task (some of this is reviewed below), no general methodology has emerged (except, perhaps, for the case of highly symmetrical specimens). Therefore, to improve the quantitative capabilities of 3D EM, we need to find ways of explicitly incorporating the contrast transfer function (CTF) of the microscope into new reconstruction algorithms that are designed from the outset to work under the high-noise conditions typical of well-preserved biological macromolecules. The two main classes of problems that we address below are CTF effects and irregularly distributed projections.

**CTF Effects.** 3D reconstruction in EM is based on various assumptions, the most important being that the negatives of the EM images are approximately the projections (line integrals) of the original volume, that is:

\[ w_i = P_i x, \]  

where \( w_i \) is the negative of the \( i \)th experimental image, \( P_i \) is a projection operator, and \( x \) is the specimen of interest. Because \( w_i \) and \( -w_i \) are negatives of each other, either can be considered to be the experimental image. Although at this point it may appear more natural to give this name to \( w_i \), the soon-to-be-discussed model of image formation caused us to adopt the opposite convention. There are many well-established techniques that will correctly estimate \( x \) from the \( w_i \), provided that Eq. (1) is a good model of the experimental data collection (Herman, 1980). Unfortunately, especially at high resolution, the negatives of EM images are not faithful representations of the specimen’s ideal projections. Therefore, a volume reconstructed based on Eq. (1) is only a rough approximation of the original specimen. A more accurate (but still somewhat simplified) model of EM imaging is

\[ y_i = h_i \otimes P_i x. \]  

In words, the recorded image \( y_i \) is similar to what we would obtain if we convolved the ideal projection image \( P_i x \) with a point spread function (psf) \( h_i \) (Hawkes and Kasper, 1996), see Fig. 1; some actual psf are plotted in the left panel of Fig. 8. Note that because the central peak of the psf is negative, a zeroth-order approximation of the projection \( P_i x \) is the negative of the \( y_i \) of Eq. (2). The Fourier transform \( H_i \) of the psf is referred to as the CTF. Two types of approaches have been proposed to overcome the CTF effects.

Approaches of the first type estimate \( w_i \) from \( y_i \); i.e., they attempt to calculate the ideal data and then solve Eq. (1). They often work in Fourier space. If we denote the Fourier transform of \( y_i \) and \( w_i \) by \( Y_i \) and \( W_i \), then

\[ Y_i = H_i W_i, \]  

Therefore, it is tempting to calculate \( W_i \) at a frequency \( k \) as

\[ W_i(k) = \frac{Y_i(k)}{H_i(k)}. \]  

However, in practice, this approach does not give useful results because the \( Y_i \) are corrupted by noise that is greatly amplified in the frequency regions in which the CTF is close to zero (and it is useless where it is actually zero).

Another example of the first approach is Wiener filtering (Marks, 1996), which can be described as a careful division by the CTF to avoid noise amplification. Wiener filters amplify the data at frequencies where the CTF has moderate to large values and suppress it at frequencies where the CTF has small values. Although this method improves the quality of the reconstruction, Wiener filters are often criticized for their limited recovery of resolution.

Approaches of the second type try to solve Eq. (2) directly (Zhu et al., 1997a; Skoglund et al., 1996). These approaches calculate a volume \( x \) that minimizes the difference between the left and right sides of Eq. (2) either in a least squares sense (Zhu et al., 1997a) or with a maximum entropy criterion for \( x \) (Skoglund et al., 1996).

**Irregularly Distributed Projections.** For single-particle 3D reconstruction, the development of angular assignment techniques (Cheng et al., 1994; Penczek et al., 1994; Radermacher et al., 1994; van Heel, 1987) has allowed the collection of large amounts of cryoelectron microscope images of unsilted specimens. However, when particles have a preferential orientation within the ice layer, one type of view is oversampled. Concerns have been raised about 3D reconstruction algorithm performance when the angular distribution of the projections is highly uneven (Boisset et al., 1998, describe the problem in conjunction with a simultaneous iterative reconstruction technique [SIRT]). Hence, there is a need for reconstruction algorithms capable of efficacious performance even for irregularly distributed projection directions.
B. Incorporation of Prior Knowledge.

Incorporation of Information from Other Microscopies. The variety of microscopic imaging techniques in molecular biology for producing 3D representations results in a proliferation of volume data for the same object, but almost no work has been invested in interrelating them. The current practice of 3D reconstruction is not—as it should be—a process that integrates all the information available for a given specimen; typically, it is a self-contained process of recovering volume information from a single data collection technique.

Our first attempt at incorporating prior information into EM reconstruction was to apply projections onto convex sets (POCS; Carazo and Carrascosa, 1987a,b) using the spatial constraint that the specimen must be contained inside a sphere of known radius. This simple constraint does not significantly improve the results. POCS has been applied to EM by Akey and Radermacher (1993) and by Radermacher et al. (1994). A similar approach to the 3D reconstruction of biological macromolecules using X-ray crystallography was taken by Stroud and Agard (1979). Attempts have been made to incorporate high-resolution surface reliefs produced by AFM into 3D EM (Müller et al., 1997), but only as a postprocessing step after reconstruction. We expect that incorporation of information from other microscopies into the reconstruction process itself will result in significantly more informative reconstructed volumes.

Recent advances in the field of AFM, as well as in high-resolution metal shadowing (Smith and Kistler, 1977, 1980), have produced impressive results for a number of specimens: for instance, subnanometer resolution of the surface relief has been achieved for the δ-29 connector (Müller et al., 1997), aquaporin-1 (Heymann et al., 1998), and other membrane proteins (Cheong et al., 1993; Fotiadis et al., 1998). Thus, we are now at a point where excellent structural data are being produced by AFM and 3D EM on the same specimens (Valpuesta et al., 1999; Walz et al., 1997). However, there is no methodology for quantitatively uniting information from such distinct sources.

Incorporation of Crystal Properties. The analysis of biological material by transmission EM (TEM) suffers from the sensitivity of such material to electron radiation. To control the damage, the electron dose is kept low, which results in a bad signal-to-noise ratio in the projections. If the object is composed of subunits that are arranged in a regular manner forming a 2D crystal, image processing of the data recorded by the microscope can alleviate this problem. The application field in which the specimens are arranged as 2D crystals is usually referred to as electron crystallography. Fourier transformation is usually employed (Henderson et al., 1986; Stewart, 1988), due to the fact that it often enables the data to be manipulated more conveniently. In current practice, the 3D reconstruction itself is performed essentially in Fourier space. We propose to perform 3D reconstruction of 2D crystals in real space. There are two reasons for our belief that an iterative method in real space may be more efficacious. First, iterative real space methods (such as algebraic reconstruction techniques [ART]) are a more convenient framework to incorporate the prior information from other microscopies (Müller et al., 1997; Smith and Kistler, 1977, 1980) and from an adequate modeling of both the sample (Chan et al., 1998) and the imaging device (Amos et al., 1982). Second, iterative real space methods avoid the interpolation in Fourier domain that may generate artifacts.

C. Display and Analysis of Reconstructions. The result of a 3D EM process is an estimate of the Coulomb potential of a biological macromolecule, typically at a medium resolution (between around 6–20 Å). As a consequence of this limited resolution, it is generally impossible to perform a task such as tracing a polypeptide chain. Therefore, the problem of interpretation is quite different from that in X-ray diffraction, where the experimentally obtained 3D electron density is treated by fitting the individual electron densities of each of the amino acids to the reconstructed general electron density. Due to our limited resolution, we cannot presume that the 3D potential can be similarly decomposed. Therefore, the important task of accurately differentiating the specimen from its background necessitates the development of other types of approaches.

Accurate Visual Representation of Biological Structures. A reconstructed volume \( x \) is corrupted by noise (resulting from the noise in the data collection) and by errors inherent in the reconstruction algorithm and related preprocessing steps. As a result, \( x \) is typically an imprecise approximation of the macromolecular volume and should be further processed to obtain the 3D structure of the macromolecule. The current procedure for extracting the macromolecular representation from the volume is based on thresholding, which assumes that there is a fixed threshold such that exactly those grid points at which the value of \( x \) is greater than the threshold belong to the macromolecule. Therefore, selecting an appropriate threshold can, in principle, separate the macromolecule from its background. The currently used techniques for the selection of the threshold are visual inspection, volume histogram inspection, and volume comparison (Frank, 1996; Perkins et al., 1997). In the histogram inspection approach, a histogram is produced by plotting, for consecutive ranges of gray values, the number of grid points that have a gray value in the range. Because the volume is occupied by macromolecular matter and its background, the resulting histogram should be bimodal. Furthermore, a threshold can be found which best separates the gray values corresponding to the macromolecular peak from the gray values corresponding to the background peak. Volume comparison selects a threshold \( t \) such that if \( n \) is the number of grid points at which the value of \( x \) is greater than \( t \) and \( \delta \) is the sampling rate, then \( n \delta^3 \) matches the known molecular volume. Based on the threshold volume, a so-called isosurface is produced, whose intended property is that it separates the macromolecular matter from the background. The production of the isosurface can be done by a number of standard imaging techniques such as boundary tracking (Artzy et al., 1981) or the marching cubes algorithm (Miguet and Nicod, 1997). In any case, such an isosurface consists of many small polygons (commonly triangles or quadrilaterals) that are then used by another piece of standard software to generate the macromolecular images on the computer screen. We propose approaches for visual representation that do not involve such piecewise planar surfaces and hence are likely to produce more accurate renderings of the biological boundaries.

In spite of its widespread use, it is known that segmenting by thresholding is often not the best approach because the structural components in a volume can overlap in their gray values. Therefore, they are not defined by a range of gray values (Herman, 1998b; Pal and Majumder, 1997; Udupa and Herman, 2000; Udupa and Samarasekera, 1996). This complication leads us to consider that the methods described above can be improved by using other segmentation approaches to extract the objects from the volume. There are several such approaches: physics-based methods, artificial intelligence, statistical methods, and methods based on fuzzy sets. We outline novel variants of the current statistical and fuzzy approaches (Carvalho et al., 1999; Chellappa and Jain, 1993; Geman and Ge-
man, 1984; Pal and Majumder, 1997; Udupa and Samarasekera, 1996).

III. SOME FOREGOING DEVELOPMENTS

In this section, we survey our prior work that is relevant to the new research described in this paper.

A. Image Representation Using Blobs. We have been investigating an alternative way of representing the unknown volume distribution, as distinct from the conventional representation using voxels arranged on a simple cubic grid. The alternative representation uses elements called blobs that can be arranged on various grids (simple cubic, body-centered cubic [bcc], face-centered cubic [fcc] Herman, 1998b; Matej and Lewitt, 1995).

Blobs have spherical symmetry and are spatially limited. The basis function associated with a blob has value one at the center of the blob and is a bell-shaped function of the radius from the center, going to zero smoothly at the boundary of the blob. A blob overlaps with its neighbors on the grid. On the other hand, the basis function associated with a voxel has value one inside its cube-shaped region and value zero outside, with a discontinuity at the boundary.

The blob’s independence with respect to orientation and direction, and its continuity and smoothness, make it well suited for representing natural structures of all physical sizes. The potential advantages of blobs for iterative reconstruction and for display of the reconstructed volume are described by Lewitt (1990). Examples of the blob basis function for various choices of its parameters are given by Lewitt (1992), together with examples of its projection (which is easily evaluated using a simple formula) and its Fourier transform (which is close to zero at frequencies above a certain bandlimit). Our subsequent work (described below) confirms the predictions by Lewitt (1990, 1992) concerning the advantages of blobs.

B. Algebraic Reconstruction Techniques. ART are iterative procedures for solving systems of linear equations. They have been used in tomography to recover objects from their projections. Informally, ART can be described as an iterative algorithm that starts with a initial trial volume \( x^{(0)} \) and iteratively modifies it depending on how similar are the measured projection data and the corresponding computer-generated projection of the current trial volume (Fig. 2).

Herman (1998a), provides an up-to-date discussion of ART. Mathematically, the algebraic reconstruction techniques are series expansion methods (Herman, 1980); i.e., they assume that the volume \( x \) can be approximated by a linear combination of a finite set of known basis functions \( b_j \),

\[
x(r, \theta, \phi) = \sum_{j=1}^{J} c_j b_j(r, \theta, \phi),
\]

and the task of the algorithm is to estimate the unknown coefficients \( c_j \).

Variants of ART in which the basis functions \( b_j \) are blobs have been tested for a variety of situations (Marabini et al., 1997, 1998, 1999; Matej et al., 1994; Matej and Lewitt, 1995). Our conclusion is that ART with blobs produces high-quality reconstructions and is superior (for a large number of tasks) to other algorithms used in the field, such as weighted backprojection (WBP) or SIRT. Therefore, we will use ART with blobs as the core method for the developments described below.

C. Iterative Data Refinement. IDR methods estimate the data that would have been collected by an idealized device (in our case, an electron microscope that collects line integrals) from data that were collected by an actual measuring device (in our case, an electron microscope with CTF blurring) as is described in Figure 3 (Censor et al., 1985). IDR has been successfully applied, among other things, to deblurring the experimental data in magnetic resonance imaging (Herman and Ro, 1990).

D. Priors. Reconstruction of piecewise smooth images can be significantly improved by Bayesian estimation using specifically designed Gibbs priors that incorporate not only the information about the smoothness within the regions, but also about the nature of the borders that exist between the regions (Chellappa and Jain, 1993; Geman and Geman, 1984). We have studied an approach in which the borders are modeled by directly using (in the energy function of the Gibbs distribution) specific cliques of pixels that represent border elements and have shown that it provides particularly high-quality reconstructions (Chan et al., 1997, 1998).

The effectiveness of using priors for EM has been reported by Carazo et al. (1999) on some segmentation experiments; one of these is reproduced in Figure 4. The segmentations were done using a prior (which encourages the formation of homogeneous regions) that was proposed for positron emission tomography (PET) by Chan et al. (1997, 1998).

E. Discrete Tomography. DT uses an alternative type of prior information, i.e., the knowledge that the values of the volume to be reconstructed must come from a known finite set (Herman and Kuba, 1999). Most specimens investigated by cryoelectron microscopy contain essentially two (protein and ice) or three (protein, ice, and nucleic acid) components. To assess whether or not this fact may translate into the applicability of discrete tomography to EM, we have designed and performed a proof-of-concept test aimed at measuring how the biochemical and biophysical differences among ice, protein, and nucleic acid are reflected in the voxel values of a
volume at a given resolution. To this end, we generated 3D volumes made of ice only, protein only, and nucleic acid only. From these volumes, sampled at different resolutions, we calculated histograms of values (Carazo et al., 1999). Figure 5 shows that at 2.5 Å, the histograms are only slightly separated, but at 7.5 Å resolution they present negligible overlapping. We discuss below how such histograms can be used to improve 3D EM.

F. Fuzzy Segmentation. We call a sequence of pixels a chain; its links are the pairs of consecutive pixels. In fuzzy segmentation, the strength of any link is automatically defined based on statistical properties of the links within regions identified by the user as belonging to the object of interest. The strength of a chain is the strength of its weakest link. The fuzzy connectedness between any pair of pixels is the strength of the strongest chain between them. The fuzzy object containing a given seed pixel at a particular threshold is the set of all those pixels whose fuzzy connectedness to the seed pixel exceeds or equals the threshold (Udupa and Samarasekera, 1996). Carvalho et al. (1999) specified an efficient algorithm for finding such objects in an image and demonstrated its efficacious performance on images corrupted by random noise and/or shading.

G. Evaluation Methodology. Our approach to evaluating image reconstruction algorithms starts with a specification of the task for which the image is to be used and then defines a figure of merit (FOM) that determines quantitatively how helpful the image is, and hence the algorithm, for performing that task. The task-specific FOM is computed for each image. Based on the FOMs for all the images produced by two different techniques, we can calculate not only the statistical significance at which we can reject the null hypothesis that the methods are equally helpful for solving a particular task in favor of the alternative hypothesis that the method with the higher average FOM is more helpful for solving that task, but also the relevance of the improvement (Furuie et al., 1994; Matej et al., 1994, 1996).

H. Conical Tilt Geometry. This is a commonly used method of data collection in many applications of 3D EM to biological parti-
cles. Marabini et al. (1998) proposed the use of ART with blobs and compared its performance with WBP, the standard technique in the field, by applying the evaluation methodology that assigns a level of statistical significance to the claim of superiority of one algorithm over the other for a particular task. The conclusion we reached is that ART with blobs produces high-quality results and is superior to WBP in recovering features along the Z direction.

Marabini et al. (1999) gave a more thorough analysis of these algorithms. In particular, they answered two additional questions: is the influence of the first-processed projections significantly different from those that are processed last and does the position in the volume of a feature influence the quality of its reconstruction? The answer is no in both cases.

I. Electron Tomography. ET collects its data by repeatedly tilting a specimen along a single (or double) axis (rather than tilting once multiple specimens, as with conical tilt geometry). Consequently, the number of available projections is typically two orders of magnitude smaller than for conical tilt geometry (and the signal-to-noise ratio is also lower). Marabini et al. (1997) introduced some highly realistic phantoms that are no longer defined as a set of geometric figures (as in previous work), but as an ensemble of atoms—a simulation of a PDB file (Bernstein et al., 1977). The projections are generated by a virtual electron microscope in which the individual interactions between the incident wave of electrons and the sample are calculated using quantum mechanics, taking into account the limitations and instabilities of real microscopes. The results again indicate that ART outperforms WBP for a number of tasks, in particular, for recovering features along the Z direction (Fig. 6; note also the false central feature in the WBP reconstruction).

J. Irregularly Distributed Projections. Sorzano et al. (1998) provided a comparison between different 3D reconstruction algorithms (WBP, SIRT, and ART with blobs) for the case of irregularly distributed projections. We reproduce the reconstructions of the giant hemoglobin of Lumbricus terrestris (de Haas et al., 1997) from 3,099 projections of 90 × 90 pixels obtained without tilting the specimen support. The results in Figure 7 suggest that ART with blobs outperforms the other two reconstruction methods.

IV. INCORPORATION OF IMAGE FORMATION MODELS

A. CTF. We propose a number of ways of incorporating the CTF into the 3D EM reconstruction procedure.

The first way is to incorporate the CTF blurring into the system of equations from which the coefficients $c_j$ of Eq. (5) are to be estimated. The currently implemented ART with blobs procedure calculates a projection of the current trial volume (Fig. 2) by making use of the formula for the projection of a single blob (called a footprint) and creating a superimposition of copies of this footprint (shifted and weighted corresponding to the locations and the current coefficients of the blobs making up the volume). To incorporate CTF blurring, we need to recalculate the footprint so that it is a blob projection after the CTF has been taken into account. Because this can be done by numerical integration, we can use the current ART with blobs, but with the new footprint replacing the currently implemented one.

There are two difficulties with this. One is that the support (the region outside which the footprint is guaranteed to be zero valued) of the new footprint may be much larger than that of the old one, which would make the computations much more expensive. The other is that the new footprint function may not have all the favorable properties of the old one, which may make the process ill conditioned. Both of these problems can be, in principle, overcome.

Figure 6. Surface-rendering representation of (a) the original phantom, (b) reconstruction using ART, and (c) reconstruction using WBP.

Figure 7. 3D reconstruction (side views) of the Lumbricus terrestris extracellular hemoglobin performed using (a) SIRT (using the parameters suggested by de Haas et al., 1997), (b) ART, (c) WBP, all of them from the same unevenly distributed set of projections, (d) reconstruction of the same from an evenly distributed set of projections (for this case, the results obtained with the three methods are quite similar, here only that for SIRT is shown).
by choosing from the three-parameter family of blob functions ones that will result in footprints with desirable properties. A preliminary experiment (right panel of Fig. 8) suggests that, for the range of typical defocus, a footprint of about 28 Å is needed to properly take into account the CTF effects.

Reconstruction procedures that incorporate the CTF and involve blobs, or other volume basis functions, need to ensure that these basis functions (and their spacing on the grid) are compatible with the CTF. For example, the discretization of the volume should have a resolution that is compatible with the expected resolution of the defocused data. Following this idea to a more detailed level, one can argue that it would be desirable for the basis function of the volume to have a Fourier transform that goes to zero at the same frequencies where the CTF goes to zero. Our second proposed method of CTF incorporation achieves this in a simple manner by using the volume psf as the basis function for the volume, instead of using voxel or blob basis functions. Choosing the volume psf as the basis function should improve the stability of the method, because it would not attempt to estimate missing components at the zeros of the CTF.

In general, iterative reconstruction procedures involve fitting the footprint of the basis function to the micrograph data, so this specific method would involve fitting the micrograph psf to the data, because the footprint (line integral projection) of the volume psf is the micrograph psf $h_i$. The reconstruction procedure can make use of the present implementation of the ART algorithm, with an extended footprint table containing the 2D psf (2D inverse transform of the CTF) instead of the presently stored projection of the blob, or the 2D psf convolved with the projection of the blob, as proposed above. Preliminary calculations for typical choices of the parameters (Fig. 8) show that the footprint size is similar for the 2D psf itself and for the 2D psf convolved with the projection of the blob, and that the functions themselves have the same general shape. A consequence of the extended footprint is that the matrix corresponding to the system of linear equations will have more nonzero elements, so the forward projection and backprojection operations of the algorithm will require more computation, compared to the case when the footprint does not include the CTF model. We expect that a further increase in footprint size will be needed to model the CTF of a high-resolution microscope having a field-emission gun (FEG) as a beam source. However, we expect that the amount of computation will remain feasible.

When the 3D psf is chosen as the basis function, the continuous volume represented by the sum of the basis functions is the blurred volume, which the iterative algorithm tries to make consistent with the blurred data by adjusting the coefficients of the basis functions centered at the grid points. For this choice of basis function, each coefficient can be interpreted as an estimate of the value of the unblurred object at the corresponding grid point in 3D space. We get the deblurred volume at sample points on the grid without any extra work, by just looking at the coefficients produced by the algorithm. The object is being sampled at a rate that depends on the spacing of the grid points and so, as always, the grid spacing must be chosen appropriately according to the spectrum of the object and the resolution of the data.

This formulation allows the psf to be noncircular, which enables us to model astigmatism. It also allows the psf to be different in different parts of the volume, which enables us to model defocus variation resulting from the tilt of the specimen support and defocus variation as the electrons travel through a thick specimen. In principle, each grid point in the volume could have its own psf, which would just make the forward projection calculations more complicated.

The algorithm needs some regularization to constrain the solution, because the deblurred volume is represented by a set of sample points. This is in contrast to the algorithm described earlier, which represents the deblurred volume by a superimposition of blobs, so that the solution produced by that algorithm is constrained to be a member of a space of smooth functions.

The third way of incorporating the CTF is based on the much cited algorithm of Chahine (1970), a version of which was introduced into the medical image reconstruction literature as the image space reconstruction algorithm (ISRA; Daube-Witherspoon and Muehllehner, 1986; De Pierro, 1987). The underlying idea is the following. The superimposition of ordinary (without CTF) ART backprojections (Fig. 2) of each of the measured projections gives us a set of blob coefficients $d_j$, $1 \leq j \leq J$. One way of stating the reconstruction problem is that we are looking for a set of coefficients $c_j$ such that if $x$ is determined by Eq. (5) and the $y_i$ are determined by Eq. (2) note that these can be calculated from the $c_j$ by using the CTF-based footprint—then the superimposition of ordinary ART backprojections of the $-y$, should also give us the set of blob coefficients $d_j$, $1 \leq j \leq J$. In other words, we need to solve the system of equations $Mc = d$, in which the $J \times J$ matrix $M$ is implemented as ART projections with the CTF-based footprint followed by ordinary ART backprojections.

The implementation of Chahine’s algorithm is simplicity itself. In addition to the arrays $c$ and $d$, we need an array $e$ of the same size. The array $d$ (the backprojection of the measured data) is never changed. The array $c$ is initialized so that each of its entries is 1 and all entries are updated in each iteration. Into the array $e$ we put $Mc$ and then we update the array $c$ by replacing each $c_i$ by $c_id_i/e_i$. Note that if $c$ satisfies $Mc = d$, then the updating will not change it at all.

It is a remarkable fact that under some conditions (such as the positivity of the $d_j$ and the nonnegativity of the entries of $M$), this
algorithm has been shown (Barcilon, 1975; De Pierro, 1987) to converge to a solution of $Mc = d$, which is in fact optimal in a well-defined sense. Unfortunately, although the assumed conditions are reasonable for the applications of Barcilon (1975) and De Pierro (1987), they may well be violated by the $M$ and $d$ of the process described above. For example, it is not guaranteed in our case that the entries of $M$ are nonnegative: the CTF projections of a single blob will have both positive and negative values and there is no reason to expect that the ART backprojections of these would result in only nonnegative coefficients. We propose a three-pronged approach to overcome this difficulty: (1) look for proofs of convergence under less demanding conditions, (2) search for a choice of blob functions and implementation of backprojection process that will result in an $M$ and $d$ for which we have a proof of convergence, and (3) look for variants of Chahine’s algorithm that converge where the original version fails to do so.

Our fourth proposed way of taking care of CTF blurring is using the already-discussed method of IDR (Fig. 3). This is an “estimate $w_i$ from $y_i$” approach. It is especially appealing because it combines information coming from different images (with different CTFs) and corrects them even at the frequencies near to the zeros of the individual CTFs. Also, we emphasize here an aspect common to all our approaches: they are capable of dealing with more realistic (and complicated) models of CTF than that indicated in Eq. (2). For example, the actual device operator in Figure 3 can be replaced by one in which different parts of the object are at different defocus; this will be necessary for imaging relatively large objects at high resolution.

In discussing these four approaches, we have made the (unstated) assumption that we will be capable of modeling the actual CTF of the instrument. However, the parameters of the CTF are themselves estimated from our (noisy) electron micrographs; we now present a way of handling the uncertainty in our estimates of the CTF parameters. We refer to the middle-to-high resolution region, because current parametric models of CTF are inappropriate at low resolution.

The standard approaches to image reconstruction assume that the model (as represented by the system matrix) does not have noise or errors—it is assumed that the noise and errors are confined to the data. However, even after the most careful modeling, the actual process of data acquisition will not be described exactly by the system matrix. A more general formulation of the image reconstruction problem must allow for unknown errors in the elements of the system matrix, as well as the usual unknown errors in the data. In the numerical analysis context, this formulation is known as total least squares (TLS) and in the statistical context it is known as errors in variables. A comprehensive survey of TLS is given by Van Huffel and Vandewalle (1991) and more recent developments in the field are described by Van Huffel (1997). Iterative reconstruction algorithms based on the TLS formulation appear to have great potential for 3D EM, but have not yet been investigated.

A specific element of the system matrix represents the contribution of a specific basis function of the volume to a specific item of the data. There are statistical uncertainties in the values of the elements of the system matrix because these values are derived from the CTF model, which is subject to errors in modeling and to errors in parameter estimation. Modeling errors arise because the CTF is an idealized description of the physical process of data acquisition. Parameter estimation errors arise because the parameters of the CTF model are determined from noisy micrographs.

The system matrix $T$ maps the vector $c$ of the (unknown) coefficients of the basis functions into a vector $p$ whose components are the measured gray values in (all of) the TEM images to be used in the reconstruction. Because $p$ contains noise and errors, typically there is no $c$ that exactly satisfies $Tc = p$; i.e., the system of equations is not consistent.

The formulation of TLS is based on the linear model $(T + E)c = p + r$, where we interpret the matrix $E$ as the smallest correction to the matrix $T$, together with the vector $r$ as the smallest correction to the data vector $p$, which will produce the consistent system of equations $(T + E)c = p + r$. The TLS estimate of $c$ is obtained by solving the TLS problem

$$\text{minimize}\|E, r\|_F \text{ subject to } (T + E)c = p + r.\quad(6)$$

Here, $[E, r]$ denotes the matrix $E$ augmented by an additional column containing the vector $r$, and $\|A\|_F$ denotes the Frobenius matrix norm $\sqrt{\sum\sum a_{ij}^2}$. For comparison, the TLS formulation reduces to the standard least squares formulation when the zero matrix is substituted for $E$, corresponding to the assumption that $T$ is known exactly.

The TLS formulation of the inverse problem has been applied, with encouraging results, in image restoration (Mesarović et al., 1995) and in light-photon transmission tomography for diagnostic radiology (Zhu et al., 1997b). An iterative method, such as the one described by Zhu et al. (1997b), is needed to process the huge and unstructured system matrices that arise in tomographic inverse problems. Although there are many significant differences between light-photon tomography and 3D EM, the reported results of simulations are very encouraging and are a strong motivation for investigating this method in the 3D EM environment. The algorithm based on TLS performed better than a standard least squares algorithm, both for random errors in the weights (i.e., elements of the system matrix) and for systematic errors in the weights. The iterative algorithm developed for light-photon tomography has some features in common with (but is not identical to) the iterative algorithms for TLS that are described by Björck (1997) in the context of numerical analysis.

The iterative method for TLS described by Zhu et al. (1997b) is an excellent starting point for investigating the potential benefits of the TLS formulation for 3D EM. Note that our use of blob basis functions in the image representation and in the system model automatically imposes a global smoothness constraint on the solution, without requiring any changes to the numerical algorithm itself. Therefore, it will be interesting to investigate the performance of the combination of blobs and the iterative method for TLS, where the blobs introduce global smoothing regularization.

The TLS formulation can be extended to include regularization. Some promising work in this direction has been done by Hansen and O’Leary (1997), but it is not clear if their methods are feasible for the very large system matrix in 3D EM. Another promising approach is the extension of the general iterative methods for inverse problems involving POCS (Carazo, 1992; Carazo and Carrascosa, 1987a,b). Most of the present methods involving POCS assume that the system model is accurate, so their formulation needs to be extended to accommodate TLS (Sharma and Trussell, 1997). There is also the intriguing possibility of using ART to solve the consistent system of equations within the TLS formulation. An analogous approach for standard least squares, with regularization, involves using ART with an augmented matrix and vector (Herman, 1980, pp. 189–192).
B. Reconstruction from Irregularly Distributed Projections. Although we are pleased with the high-quality performance of ART with blobs from irregularly distributed projections (Fig. 7), our current understanding as to why this happens is vague. We see that because ART corrects the trial volume projection by projection, it follows that if a particular projection is already more or less satisfied, then the difference that is backprojected will be small (Fig. 2). The overabundance of projections around a particular direction will not cause an overcorrection in that direction, as it may happen with a technique that uses all projections simultaneously. However, this qualitative description is not by itself useful; what we need is a theoretical understanding of the relationship for the various 3D EM reconstruction algorithms between the distribution of the projections and the resulting volume.

Related to this development is the fact that the projection directions themselves often need to be estimated from the noisy TEM images. We again have a situation in which there are random errors in some of the assumed system parameters (in this case, the calculated Euler angles of the projection directions). We expect that the incorporation of the TLS methodology into reconstruction algorithms to take care of our uncertainty in the Euler angles will result in their improved performance.

V. INCORPORATION OF PRIOR KNOWLEDGE

A. Analysis of 2D Crystals Using ART. We use the two-component column vector \( \mathbf{v} \) to denote the location in the horizontal plane through the origin and the variable \( z \) to denote the vertical height above this plane. In this notation, a 2D crystal (with translation vectors \( \mathbf{p}_1 \) and \( \mathbf{p}_2 \)) is represented by a function \( X \) of three variables, such that

\[
X\left(\frac{\mathbf{v} + \mathbf{p}_1}{z}\right) = X\left(\frac{\mathbf{v} + \mathbf{p}_2}{z}\right) = X\left(\frac{\mathbf{v}}{z}\right), \quad \text{for all points } \left(\frac{\mathbf{v}}{z}\right),
\]

and the value of \( X \) is 0 unless \( 0 \leq z \leq Z \), where \( Z \) is the maximum height of the crystal. The associated unit cell is

\[
\left\{ \left(\frac{\mathbf{u}_1 \mathbf{p}_1 + \mathbf{u}_2 \mathbf{p}_2}{z}\right) \left| 0 \leq \mathbf{u}_1 < 1, 0 \leq \mathbf{u}_2 < 1, 0 \leq z \leq Z \right. \right\}.
\]  

The projection onto a plane of a 2D crystal is also a 2D crystal (a flat one with \( Z = 0 \)). This property is retained even if we simulate CTF as in Eq. (2). The translation vectors of the projected crystal are the projections of the translation vectors of the original crystal. We could attempt to reconstruct from the measured projections only within the projected unit cell, but this would throw away one of the main advantages of collecting data using crystals: the periodic repeats contain independent random noise and the signal-to-noise ratio can be greatly increased by averaging them. This is normally done using a Fourier space method, which we now describe, because it leads to a final twist that allows us to estimate at regularly spaced sample points within the projected unit cell the averaged values of the projections.

We use \( \chi(v) \) to denote the projected value at the point \( v \) (we drop the unnecessary variable \( z \)) for a particular projection. For these calculations, the projections are processed independently of each other. We define the matrix

\[
P = \begin{pmatrix} p_1^T \\ p_2^T \end{pmatrix}.
\]
where \( p_1 \) and \( p_2 \) are the translation vectors of the projection \( y \) (which is, therefore, said to be a crystal of period \( P \)).

Using this matrix, we can define a crystal \( g \) of period \( I \) (the identity matrix), by

\[
g(v) = y(P^T v).
\]

(10)

It is a standard result that the 2D Fourier transform of such a crystal \( g \) is of the form

\[
\sum_{K=-\infty}^{\infty} \sum_{L=-\infty}^{\infty} b_{KL} \delta(\xi) \delta(\eta),
\]

(11)

where the \( b_{KL} \) are complex numbers and \( \delta(\xi) \) is the impulse function at the point \( \left( \frac{K}{I}, \frac{L}{J} \right) \). It is also well known that

\[
\delta(\xi)(PV) = \frac{1}{|P|} \delta_{P^{-1}}(\xi)(V),
\]

(12)

where \( |P| \) is the determinant of \( P \). Putting all this together, the Fourier transform of \( y \) can be approximated (by choosing an \( M \) and an \( N \) that are sufficiently large) as

\[
Y(V) = |P| \sum_{K=-\infty}^{\infty} \sum_{L=-\infty}^{\infty} b_{KL} \delta(\xi)(PV) = \sum_{K=-M}^{M} \sum_{L=-N}^{N} b_{KL} \delta_{P^{-1}}(\xi)(V).
\]

(13)

In practice (Amos et al., 1982), \( Y \) is calculated by taking the discrete Fourier transform of the measured projection \( y \). The result is a collection of impulses corrupted by noise and inaccuracies, from which the \( b_{KL} \) (and, if it is not known, the matrix \( P \)) can be estimated. Techniques have been developed mainly by the Medical Research Council (MRC) Laboratory to make these estimations (Amos et al., 1982; Crowther et al., 1996), which we will adopt for our purposes. Our contribution is what we do with the estimates. The MRC approach is to use them (from the various projections) to estimate by interpolation the 3D Fourier transform of the volume to be reconstructed on a cubic grid and then to obtain the volume by an inverse 3D discrete Fourier transform. As opposed to this, we push the mathematics further, by deriving that, for \(-M \leq k \leq M\) and \(-N \leq l \leq N\),

\[
y \left( P^T \left( \begin{array}{c}
k \\
l \\
\frac{2M + 1}{2N + 1}
\end{array} \right) \right) = \sum_{K=-M}^{M} \sum_{L=-N}^{N} b_{KL} e^{2\pi i \left( \frac{Kk}{2M + 1} + \frac{Ll}{2N + 1} \right)},
\]

(14)

which is the inverse 2D discrete Fourier transform of the array of weights in the previous formula. This provides us with the uniformly sampled averaged values within the unit cell of the projection that we need for reconstruction using ART.

We present our first results obtained by using ART with blobs for crystals. The phantom described in Figure 11(a) has been reconstructed using the traditional (MRC) algorithms based on interpolation in Fourier space (Figure 11b) and also using our new approach (Figure 11c). The reconstructions suggest that the accuracy of the traditional algorithm (at least for this particular phantom) depends on the position of the features of interest (the central sphere in the reconstruction is better defined than the spheres above and below it); this is not the case for ART with blobs. We note that these reconstructions were made from 13 projections and without taking the symmetries into consideration. A further test indicated that, at least for this unrealistic phantom, ART with blobs retains its superiority from this point of view even if we use 26 projections and make use of the fourfold symmetry.

Although eliminating the potentially troublesome interpolation in Fourier space is a desirable feature of our new approach, its main potential benefit lies in its flexibility. In addition to the already discussed capability of ART with blobs to accommodate incorporation of image formation models, one can also incorporate various kinds of prior knowledge about the volume. For example, information that the reconstructed values should lie within a certain range...
provides us with linear inequalities in the unknown coefficients. For any fixed \((r, \theta, \phi)\), the right-hand side of Eq. (5) must lie between the two limits. Also, the sum needs to be taken over only that small set of coefficients \(c_j\) for which \(b_j(r, \theta, \phi) \neq 0\). There is a version of ART to handle such inequalities (Herman, 1980, pp. 192–193; Herman, 1998a, pp. 37–38). Its output is illustrated in Figure 11(d) for the case in which the only prior information is the nonnegativity of the volume values.

The flexibility of ART opens up a large territory to be explored: the incorporation of information from other microscopies and the use of such methods to improve on the reconstructions that can be obtained from TEM data alone.

B. Priors. The Bayesian or maximum a posteriori probability (MAP) approach is built on two components: a prior distribution that for any volume tells us the probability of it occurring in the application of interest and an image formation model that for any volume and measurement data tells us the probability of those measurements arising from that volume. The principle is that we select that volume for which the product of these two probabilities is maximum.

Such methods have been used for EM only occasionally (Carazo, 1992), in spite of their demonstrated capability for solving image reconstruction problems with limitations in data collection (Chan et al., 1997, 1998). This may be due to the three-dimensionality of the problem and the lack of suitable priors. The significance for EM of using prior distributions that truly model certain global properties of the structures to be reconstructed in the Bayesian framework lies in the undoubtedly expected significant improvement in the quality of reconstructions. The introduction of prior information in the reconstruction process in EM is of particular importance due to inherent limitations of the measurement procedures.

For this, one needs probabilistic models for volumes that are likely to occur in specific areas of 3D EM. These should be truly volume-modeling distributions, in the sense that random samples from such distributions indeed exhibit the properties of the volumes to be modeled.

To develop a statistical volume model that is truly volume modeling for a class of volumes, all essential characteristics of the volumes have to be taken into account. For instance, a model of piecewise homogeneous volumes in 3D should incorporate not only the information about the smoothness within regions in the volume, but also the continuity of surface structures that exist between regions. This can be achieved using a Gibbs distribution (Geman and Geman, 1984)

\[
\Pi(x) = \frac{1}{v} e^{-H(x)},
\]

where \(x\) is a volume, \(v\) is a normalizer, and \(H(x)\) is called the energy function, which is a sum of clique potentials. Clique potentials for 2D are illustrated in Figure 12; clique potentials of this kind were used to produce the segmentation in Figure 4. Such results are promising (especially if we take into account that the prior and image formation model have not been optimized for the application), suggesting that the EM volumes can be considerably improved using MAP.

When implementing this idea for 3D, we superimpose a grid of points on that part of space that is occupied by the volume. With each grid point, we associate a voxel (which consists of all points in space that are not nearer to any other grid point; Herman, 1998b) and we consider points to be neighbors only if their voxels have points in common. The fcc grid is particularly attractive from this point of view; its voxels are rhombic dodecahedra (all isometric). With every voxel, we can associate a clique containing the grid point in that voxel and in the 12 voxels that share a face with it, all of which are at the same distance from it. Compare this with Fig. 12: the grid points in the four edge-adjacent pixels are at a different distance from the center than the grid points of the four vertex-adjacent pixels. Such properties make the fcc grid desirable for MAP processing, which (together with some additional desirable properties for analysis, to be explained below) mandates that we should develop a thorough understanding of 3D reconstruction algorithms using it. Due to its superior sampling properties, to date we have been concentrating on 3D algorithms using the bcc grid advocated by Matej and Lewitt, 1995. For whatever grid, the creation of a volume-modeling prior and an image formation model is very application dependent (Chan et al., 1997, 1998; Herman and Kuba, 1999).

As an example, MAP estimation can be based on Figure 5. Suppose that we have used ART with basis functions whose value is 1 inside a voxel of the fcc grid and 0 outside it. We would like to label each grid point as being in ice, protein, or RNA. For whatever grid, the creation of a volume-modeling prior and an image formation model is very application dependent (Chan et al., 1997, 1998; Herman and Kuba, 1999).

VI. DISPLAY AND ANALYSIS OF RECONSTRUCTIONS

A. Fuzzy Segmentation of Multiple Objects. Fuzzy connectedness has been effectively used to segment out an object in a badly

![Figure 12. Sample clique configurations for a horizontal and for a diagonal border element in 2D. The value of the clique potential increases (and so does the probability of the appearance of the configuration in the reconstructed object) whenever all a-pixels are labeled the same, all b-pixels are labeled the same but differently from a-pixels, and all x-pixels are labeled the same as either the a-pixels or the b-pixels. The a-pixels and b-pixels belong to different sides of the border.](image-url)
corrupted 2D image (Carvalho et al., 1999; Udupa and Saramasikeru, 1996). Its theory easily generalizes to pictures over arbitrary digital spaces (Herman, 1998b, Section 5.2). We implemented the approach for various 3D grids (Fig. 13).

We need to generalize the approach to simultaneous segmentation of multiple objects for two reasons. One is that sometimes there are two or more components in the volume (e.g., protein, ice, and nucleic acid). The other is that even when segmenting out a single object from its background, current fuzzy segmentation techniques do this by thresholding the fuzzy connectedness map and so an incorrect choice of the threshold results in the segmented volume not corresponding to biological reality. A way to avoid having to select a threshold is to consider the background as another object and to apply simultaneous segmentation.

For a positive integer \( M \), an \( M \)-semisegmentation of a set \( V \) is a function \( \sigma \) that maps each \( c \in V \) into an \((M+1)\)-dimensional vector \( \sigma' = (\sigma_0', \sigma_1', \ldots, \sigma_M') \), such that \( \sigma_0' \in [0, 1] \) and, for at least one \( m \) in the range \( 1 \leq m \leq M \), \( \sigma_m' = 1 \), and for all other \( m \) it is either 0 or \( \sigma_0' \). We say that \( \sigma \) is an \( M \)-segmentation if, for every \( c \), \( \sigma_0' \) is positive.

A fuzzy affinity on a set \( V \) is a function \( \psi : V^2 \rightarrow [0, 1] \). We think of \((c, d)\) as a link and of \( \psi(c, d) \) as its \( \phi \)-strength. We define a chain in \( U(\subseteq V) \) from \( c^{(0)} \) to \( c^{(K)} \) to be a sequence \((c^{(0)}, \ldots, c^{(K)})\) of points in \( U \) and the \( \phi \)-strength of this chain as the \( \phi \)-strength of its weakest link \((c^{(0)}, c^{(1)}, \ldots, c^{(K)})\), \( 1 \leq k \leq K \). We say that \( U \) is \( \psi \)-connected if for every pair of distinct points in \( U \) there is a chain in \( U \) of positive \( \psi \)-strength from the first point of the pair to the second.

If there are multiple objects to be segmented, it is reasonable that each should have its own fuzzy affinity. An \( M \)-fuzzy graph is a pair \((V, \Psi)\), where \( V \) is a nonempty finite set and \( \Psi = (\psi_1, \ldots, \psi_M) \) with \( \psi_m \) (for \( 1 \leq m \leq M \)) a fuzzy affinity such that \( V \) is \((\min_{1 \leq m \leq M} \psi_m)\)-connected. For an \( M \)-semisegmentation \( \sigma \) of \( V \) and for \( 1 \leq m \leq M \), the chain \((c^{(0)}, \ldots, c^{(K)})\) is said to be an \( \sigma m \)-chain if \( \sigma_m' > 0 \), for \( 0 \leq k \leq K \). Further, for \( U \subseteq V, W \subseteq V \), and \( c \in V \), we use \( \mu_{m, u, w}(c) \) to denote the maximal \( \psi_m \)-strength of an \( \sigma m \)-chain in \( U \) from a spel in \( W \) to \( c \).

**Theorem.** If \((V, \Psi)\) is an \( M \)-fuzzy graph and, for \( 1 \leq m \leq M \), \( V_m \) is a subset (of seed points) of \( V \) such that at least one of these subsets is nonempty, then there exists a unique \( M \)-semisegmentation (which is, in fact, an \( M \)-segmentation) \( \sigma \) of \( V \) with the following property. For every \( c \in V \), if for \( 1 \leq n \leq M \)

\[
\sigma_n'(c) = \begin{cases} 1, & \max_{d \in V}(\min(\mu_{n, u, w}(d), \psi_n(d, c))), \\ \text{otherwise}, & \end{cases}
\]

then for \( 1 \leq m \leq M \)

\[
\sigma_m' = \begin{cases} \sigma_n', & \text{if } \sigma_n' \geq \sigma_m' \text{ for } 1 \leq n \leq M, \\ 0, & \text{otherwise}. \end{cases}
\]

To see the desirability of the \( M \)-segmentation whose existence (and uniqueness) is guaranteed by this theorem, consider the following. Pick a point \( c \) in \( V \) (Fig. 14). Suppose that we know for every other point \( d \) the fuzzy membership vector \( \sigma_d \). Then, the \( \sigma_n' \) of Eq. (16) is the maximal \( \psi_n \)-strength of a chain with the following properties: its first point is a seed point in \( V_m \), its last point is \( c \), and for all intermediate points \( d \) in the chain \( \sigma_d > 0 \) (i.e., these points belong to the \( n \)th object). The objects that can most strongly “claim” to contain \( c \) are those for which \( \sigma_n' \) is maximal. This is indeed how

![Figure 13. Slice of a reconstructed volume of the SV40 large T antigen (San Martin et al., 1997) with seed pixels (a), the map for the same slice of the 3D fuzzy connectedness to the seed pixels (b), thresholded version of the connectedness map (c).](image)

![Figure 14. A point c should belong to those objects of an M-segmentation that have the strongest claim on it.](image)
things get sorted out in Eq. (17): \( \sigma_n^m \) has a positive value only for such objects.

An area where further investigation is needed is the method of determining the fuzzy affinities to be used in such an algorithm. For single-object 2D medical applications, we have found the following simple procedure efficacious. The user points at some pixels and \( V \) is formed by them and their neighbors. The program calculates the mean and standard deviation of gray values of the pixels in \( V \) and of the absolute differences between gray values of pairs of neighboring pixels in \( V \). Then \( \phi(c, d) \) is 0 if \( c \) and \( d \) are not neighbors and is otherwise calculated from the mean and the absolute difference of the gray values of \( c \) and \( d \). There is no reason to believe that this method would also be the appropriate one for multiojective segmentation of 3D EM volumes, especially in view of the following novel idea.

A reconstruction using ART with blobs calculates the coefficients \( c_j \) and this is followed by an evaluation of the values at points of the volume using Eq. (5). If these values are to be used only as input into a segmentation process, then the question arises whether we should do the segmentation on the picture defined by the grid points at which the blobs \( b_j \) are centered, each assigned as its gray value the corresponding \( c_j \). This would eliminate the interpolation of Eq. (5) and would reduce the computational burden of segmentation because Eq. (5) is typically evaluated at many more points than there are basis functions. On the other hand, because the picture of the coefficients is essentially a high-pass-filtered version of the volume, we need to reinvestigate which affinities are likely to be useful for fuzzy segmentation. We note that here again it is preferable to have the blobs placed on an fcc grid; this results in each point having 12 positive-strength links, each of the same “length.” This is likely to make affinity determination easier than if we used links of different lengths.

B. Averaging Structural Components. The functional form of many biological macromolecules is obtained by merging or joining several copies of one or several identically structured subunits. The identification and localization of such subunits is of great importance for understanding the conformational changes a macromolecule can experience and the behavior of the macromolecule when interacting with other proteins. The currently used processes for reconstructing macromolecules typically yield volumes from which it is clear that the macrostructure has repetitive subunits and yet these subunits are not sharply distinguishable. If the individual subunits could be detected, aligned, and averaged, then the signal-to-noise ratio for the subunits (and for the macromolecule) could be increased. To identify and extract subunits in reconstructed volumes, one can make use of the just-discussed fuzzy segmentation for multiple objects methodology. This approach allows us to generate exquisite renderings of surfaces (Fig. 15), but (in its current implementations) it is slower than the methods that rely on rendering a polygonal approximation of the surface. In particular, we cannot currently match the capability of the latter methods in real-time rendering of the surface from varying points of view under the interactive control of the user.

The implementation of this visualization procedure requires some attention. The most straightforward way to achieve such surface displays is by some variant of the method of “ray-casting” (Watt, 1999). This approach allows us to generate exquisite renderings of surfaces (Fig. 15), but (in its current implementations) it is slower than the methods that rely on rendering a polygonal approximation of the surface. In particular, we cannot currently match the capability of the latter methods in real-time rendering of the surface from varying points of view under the interactive control of the user. Because such a capability is quite desirable for the study of the reconstructed macromolecules, alternative approaches that retain the accuracy of rendering using Eqs. (18) and (19), but at speeds comparable with what can be done for polygonal approximations, are needed. One promising approach is based on “shell rendering” (Udupa and Odhner, 1993). In the adaptation of this to our situation, we put together \( B \) from only those basis functions that intersect it and thereby reduce the data set that represents it (and its normals) to an extent that allows real-time interactive visualization.
VII. VALIDATION OF CLAIMS OF SUPERIORITY

Claims of the superiority of a newly introduced approach over methods already in use in a field need to be validated. We have not yet done so for the methods advocated in the previous sections; in this section, we discuss our planned approach to their validation. To be convincing, such a validation should go well beyond providing anecdotal evidence based on a few experiments. We and others have been advocates of rigorous validation approaches in medical imaging, contributing both to methodological development (Furuie et al., 1994; Matej et al., 1996) and to its application (Chan et al., 1997; Kohn et al., 1991; Matej et al., 1994; Matej and Lewitt, 1995; Narayan and Herman, 1999), and to the adaptation of such approaches to 3D EM (Marabini et al., 1997, 1998, 1999).

A rigorous comparison between methods involves three different questions: What is going to be compared? (i.e., what will be the test data sets?); on which bases will the comparison be done? (i.e., what will be the FOM?); how will the comparison be performed? (i.e., what will be used for statistical analysis?).

A. Test Data Sets. We plan to use test data sets at three levels of increasing biological and EM instrumental realism. At the lowest level of realism, we use mathematically defined phantoms from which projection images (with or without incorporating a CTF) are calculated. For such data, we have a complete knowledge of the objects to be reconstructed (and of the image formation process). This level is useful for indicating basic differences between algorithms (Marabini et al., 1997, 1998, 1999).

In order to be useful, the test data sets must be based on a number of similar, but not identical, phantoms; the ability to distinguish these from each other can then be used to measure the performance of the image processing algorithms under investigation. For example, for one of the tests of the performance of algorithms for the reconstruction and for the ensuing display and analysis of 2D crystals, we will create various unit cells from four not necessarily identical structural components. A structural component is formed by two sets of at most four spheres each and a central hollow cylinder. The sets of spheres are located at opposite ends of the cylinder and around the exterior of the cylinder; we can arbitrarily name them top and bottom (Fig. 16). In each unit cell, two adjacent structural components are so near to each other that at each level only one of them can have a sphere. We design the unit cells so that exactly one of them will have a sphere. Consequently, there are $2^{16}$ possible unit cells of this type. Even though some of these define essentially the same 2D crystal (e.g., interchanging top and bottom in Fig. 16 would change the unit cell, but the resulting 2D crystals can be obtained from each other by translation), there is a sufficient number of essentially different 2D crystals of this kind that distinguishing between them based on reconstructed volumes is a challenge. This is true even if we make the further restriction that we wish some or all of the four structural components to be isometric (i.e., to be identically structured subunits) so that we can use the phantoms for validating our procedures for averaging structural components. In Figure 16, S2 can be obtained from S1 by a translation followed by a 180° rotation around $p_1$, S3 can be obtained from S2 by a translation, and S4 can be obtained from S1 by a translation.

For such data sets to be useful for distinguishing between imaging algorithms, the task to be performed must be hard (otherwise all algorithms will succeed in performing it), but not impossibly difficult (otherwise none will succeed). For example, if in the phantoms just described the spheres are large as compared to the resolution of the instrument but small as compared to the gap between the cylinders, then the task of assigning the individual spheres to the various structural components will be too easy. On the other hand, if the spheres are so large that they practically touch the adjacent structural component, then it will be unreasonable to demand the identification of the structural component containing a particular sphere. Therefore, the exact design of the phantoms of a general class requires careful preliminary work, which will have to be repeated for different instrumental parameters (e.g., different CTF, different noise level).

At the next level of realism, phantoms will be produced from structures of biological specimens that have been solved at atomic resolution (and, typically, deposited in the PDB) and the imaging process will be simulated using the precise quantum mechanical approach of Dinges and Rose (1995). Marabini et al. (1997) used such an approach for a phantom that was produced by R. Schröder by cutting spherical pieces from the structure of G-actin and then simulating, at the atomic level, the embedding of these spheres into an aqueous environment (Fig. 6). At this level of realism, we consider the characteristic properties of the biological structures from the very beginning. This is especially important at high resolution, because it opens the possibility to define phantoms with specific folds (defined sequences of secondary structure that occur in proteins) and to study how different algorithms reconstruct this fold depending on a number of circum-
stances, such as the data collection geometry that is being simulated. Also, there is no assumption of any CTF model. The simulated images are obtained instead by a highly realistic process.

We plan to complement this with test data sets of ultimate realism, consisting of experimental images of specimens whose structure has been (or will have been) obtained at atomic resolution by other means. Consider, for instance, the φ-29 connectors, for which we already have TEM images at better than 10 Å (Valpuesta et al., 1999). The atomic-resolution structure of the connector is likely to be solved in the next couple of years (Gaussch et al., 1998; Tao et al., 1998). Obviously, once the atomic structure of the φ-29 connector is known, we will be able to ascertain the relative quality of the 3D EM reconstructions according to the different FOMs that measure the similarity of the reconstructed volume to the known structure. In the case of the φ-29 connector, we also have high-resolution surface reliefs of the crystal obtained by AFM (Müller et al., 1997), and so we can validate the usefulness of introducing such topographic information into the 3D EM reconstruction process. Additionally, in the laboratories of A. Engel and H. Gross (personal communications), a number of (membrane) proteins are being studied by a combination of techniques. These will provide test data sets that we can use for comparison purposes.

As for noncrystalline specimens, we plan to use the hexameric helicase DnAB of *Escherichia coli*. We already have a low-resolution cryoelectron microscopy map of this enzyme (San Martin et al., 1998), on which the N-terminal region has been localized by immunomicroscopy (our result, manuscript in preparation). The atomic resolution structure of the N-terminal fragment has been solved by NMR spectroscopy (Weigelt et al., 1999). We expect to dock it into a higher-resolution version of our cryoelectron microscopy reconstruction shortly. As for the carboxyl domain, the structure of the RepA protein of plasmid RSFI010 has been solved (W. Saenger, personal communication), and the structure of the helicase of bacteriophage T7 has been reported by Sawaya et al. (1999). Considering all these developments, we expect that in the not too distant future, the atomic resolution structure of DnAB will either become available or it will be possible to model it accurately from the 3D EM and NMR work on this enzyme and the X-ray work on related enzymes. At that point, we shall be able to make comparisons between different 3D EM reconstructions by measuring their similarity to the atomic model.

B. FOMs. FOMs are numerical evaluators of specific characteristics of reconstructed volumes that are indicative of the efficacy of the volumes for finding the answer to some biologically relevant questions. For example, Marabini et al. (1998) observed that the ART with blobs and the WBP reconstructions of the DnAB helicase from conical tilt TEM data differed in the height of the reconstructed particle. To decide which method is more trustworthy from this specific point of view, an FOM was designed to measure the vertical resolution of the combined TEM and reconstruction processes. Using this FOM, we were able to demonstrate the specific superiority of ART with blobs.

The previous paragraph is representative of our experience: FOMs should be designed so that they are capable of measuring the performance of two (or more) imaging algorithms from those points of view in which their output differs in a biologically relevant manner. It is, therefore, inappropriate to try to decide ahead of time the list of FOMs that will be developed; they should be designed in response to observed differences.

As an example, consider the situation depicted in Figure 16, without any restrictions on the unit cells; thus, there are $2^{16}$ possible unit cells. After we simulate the TEM image formation process, reconstruct by an image reconstruction algorithm, and apply some way of deciding for each of the 16 spheres to which of the two possible cylinders it is attached, we can be correct for 0–16 of our spheres. We can then say that the FOM of that particular reconstruction is the number of correct decisions divided by 16, which is a value between 0 (total disaster) and 1 (perfect performance).

C. Statistical Analysis. We proposed a way of assigning statistical significance to an improvement (according to an FOM) of a new image processing algorithm over an old one (Furuie et al., 1994) and a way of quantitating the relevance of the improvement (Matej et al., 1996). We provide an abbreviated discussion of this approach, using the FOM of the previous paragraph.

Suppose that we wish to validate the claim that a newly proposed ART is superior (according to this FOM) to a currently used method of the MRC package (Crowther et al., 1996). To do this, we would randomly pick a large number (e.g., 100) of phantoms of the type depicted in Figure 16, simulate the TEM image generation, and reconstruct from the projections by both algorithms. We would apply our display and analysis software to each reconstructed volume to obtain its FOM. If the average FOM for ART is lower than that for the MRC method, then the claim of superiority is immediately rejected. Otherwise, for each projection data set, we calculate the FOM for ART minus the FOM for MRC. The null hypothesis that there is no difference in performance between the two methods (as compared to the alternative hypothesis that ART is superior) translates into testing the null hypothesis that the sum of the differences in the FOMs is from a zero-mean normal distribution whose variance is the sum of the squares of the differences. For details see Furuie et al. (1994). They also recommend the use of the $t$-test for paired data in the case where the number of projection data sets is small. In either case, the statistical validity of using such tests depends on certain distributions being normal. It has been our experience in all cases where we applied this methodology that the normality assumption is valid (Chan et al., 1997; Matej et al., 1994; Narayan and Herman, 1999); however, this will have to be retested for each new application. Once this translation is made, the assignment of statistical significance to the claim of superiority is routine. In practice, we often compare the merits of multiple methods relative to each other, in which case there is a danger of committing a type I error of wrongly rejecting a true hypothesis. We will be using the method suggested by Hommel (1988) to prevent this, as we have done previously (Narayan and Herman, 1999). Finally, the relevance of the improvement is calculated as the percent of imperfection in the performance of the MRC method (defined as 1 minus the average FOM for the MRC method) that is removed by using ART instead (Matej et al., 1996).

VIII. SUMMARY

We have discussed three potential ways of improving 3D EM: A. incorporation of realistic image formation models into new reconstruction methods; B. incorporation of knowledge regarding the specimen obtained by means other than EM; and C. improving the rendering and the analysis of the reconstructed volumes. We have also outlined an approach to validating claims of superiority of the proposed methods over those used in current practice.
ACKNOWLEDGMENTS

The work of the authors is supported by NIH grants HL28438 and CA54356, NSF grant DMS9612077, NATO grant CRG960070, grant BIO980761 from the Comisión Interministerial de Ciencia y Tecnología, a fellowship to R. Marabini from the Spanish Ministry of Education and Science, a scholarship to E. Garduño from the Consejo Nacional de Ciencia y Tecnología, and a grant to S. Matej from the Whitaker Foundation.

REFERENCES


P.R. Smith and J. Kistler, Surface reliefs computed from micrographs of heavy metal-shadowed specimens, J Ultrastruct Res 61 (1977), 124–133.


S. Van Huffel (Editor), Recent advances in total least squares techniques and errors-in-variables modeling, SIAM, Philadelphia, 1997.


