JOURNAL OF STRUCTURAL BIOLOGY **120,** 146–157 (1997)
ARTICLE NO. SB973901

Identification of Spherical Virus Particles in Digitized Images of Entire Electron Micrographs

Ioana M. Boier Martin

Department of Computer Science, Indiana University, South Bend, Indiana 46634

Dan C. Marinescu and Robert E. Lynch

Department of Computer Sciences, Purdue University, West Lafayette, Indiana 47907

and

Timothy S. Baker[1]

Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907

Received March 7, 1997, and in revised form June 26, 1997

New methods are described that should facilitate high-resolution (5–10 Å) image reconstructions from low-dose, low-contrast electron micrographs of frozen-hydrated specimens and processing of large, digital images produced by new imaging devices and modern electron microscopes. Existing techniques for automatic selection of images of individual biological macromolecules from electron micrographs are inefficient or unreliable. We describe the Crosspoint method (CP), which produces good quality solutions with relatively small miss rates and few false hits, and an extension of this method along with a procedure for refining its solution. Two algorithms for processing large images, one based on image subsampling, the other on image decomposition, are described. A large image is first compressed (e.g., by subsampling) and the CP method is applied to the compressed image to produce an initial solution. The information gathered at this stage is used to cut the original image into subimages and then to refine the particle coordinates in each subimage. An interactive environment for experimenting with particle identification methods is described. © 1997 Academic Press

1 INTRODUCTION

An important goal of transmission electron microscopy is to reveal the three-dimensional structure of the specimen under study. From electron micrographs which contain many different projections of identical macromolecules (e.g., virus particles), it is possible to produce a spatial model of the structure of an average particle (see, for example, (*4*)).

The three-dimensional (3D) model of a specimen is normally represented as a density function sampled at the points of a regular grid. The images of individual particles in electron micrographs are approximate projections of the specimen in the direction of the electron beam. The problem of determining the specimen structure from the micrographs is equivalent to the problem of reconstructing a density distribution from its projections. Fourier theory provides a simple approach to finding the 3D structure of an object from its projections. The *Projection Theorem* (*6*) connects the Fourier transform of the object with the transforms of its projections.

The basic steps of the 3D reconstruction process begin with the *selection* (boxing) of individual particle images from a number of digitized micrographs (Fig. 1). These projections of the virus particles constitute the different views used to fill in the 3D Fourier transform of the specimen. The number of such views depends on the desired resolution of the final structure and on the particle size. Next, the orientations of the specimen that give rise to these projections must be determined (*9*). The best results are often obtained in the case of highly symmetrical particles such as icosahedral viruses because the high symmetry leads to redundancies in the Fourier transform data and this, in turn, aids in the orientation search process. The 3D Fourier transform of the particle is calculated from experimental values in central sections. The values of the 3D transform

[1] To whom correspondence should be addressed. Fax: (765) 496-1189. E-mail: tsb@bragg.bio.purdue.edu.

1047-8477/97 $25.00

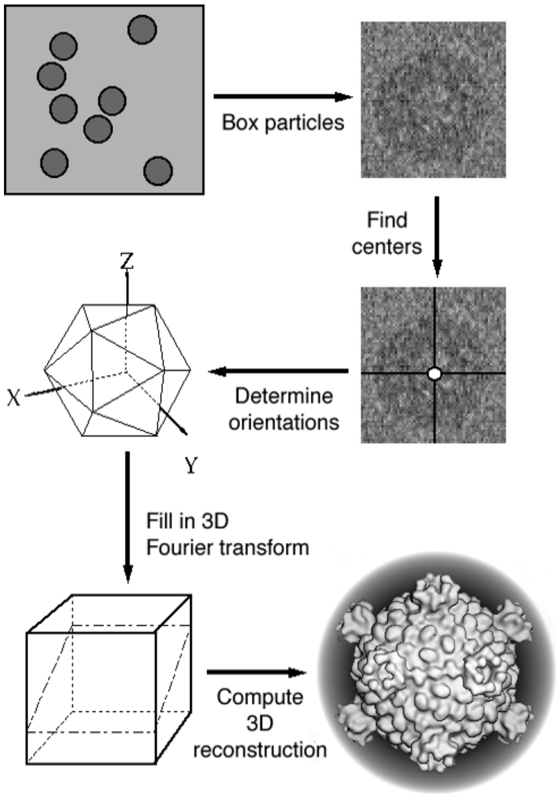


FIG. 1. Schematic representation of the steps in a three-dimensional reconstruction of a spherical virus particle from electron micrographs.

must be sampled at the points of a 3D regular grid and this requires interpolation methods (*14,18*). The last step is to compute the electron density function from the 3D Fourier transform by an inverse Fourier transformation.

Boxing, the first step in the 3D reconstruction procedure, is generally performed by a manual selection process. Because this selection procedure can be tedious, most low-resolution reconstructions (e.g., 20 Å) of relatively small virus particles have been computed from fewer than 100 particle images. It was estimated that approximately 2000 particle images are necessary for the reconstruction of a virus with a diameter of 1000 Å at 10 Å resolution (*20*), and recent results at 7–9 Å resolution with hepatitis B virus capsids (*3,5*) have confirmed this estimate. Hence, manual boxing methods are becoming impractical. The need for computer-aided particle detection methods provides the motivation for our work.

At high magnification, noise in electron micrographs of unstained, frozen-hydrated macromolecules is unavoidable (*14*) and makes automatic detection of particle positions a challenging task. Variability in the background support film of the specimen sample and radiation damage are two major sources of noise. Background variations in a micrograph can be enhanced in the digitized image by use of a color look-up table (Fig. 2). Radiation damage is the consequence of the exposure of the specimen to the electron beam required to produce high-magnification images. Limited exposure is used to maximize specimen preservation, but the result is a low-contrast image. A typical low-contrast micrograph and a histogram of the density values show that gray levels in the image are concentrated in a very narrow range, as discussed in section 2.3 and illustrated in Figs. 5a and 5c.

An ideal automatic particle selection method must produce a reliable solution and be computationally efficient. For high-resolution reconstruction work it is necessary to analyze large numbers of micrographs at speeds comparable to the data acquisition rates. New input devices such as modern scanning microdensitometers and CCD (charge-coupled device) detectors routinely allow frames consisting of 6000 × 6000 or more pixels to be collected within a time frame of minutes or less.

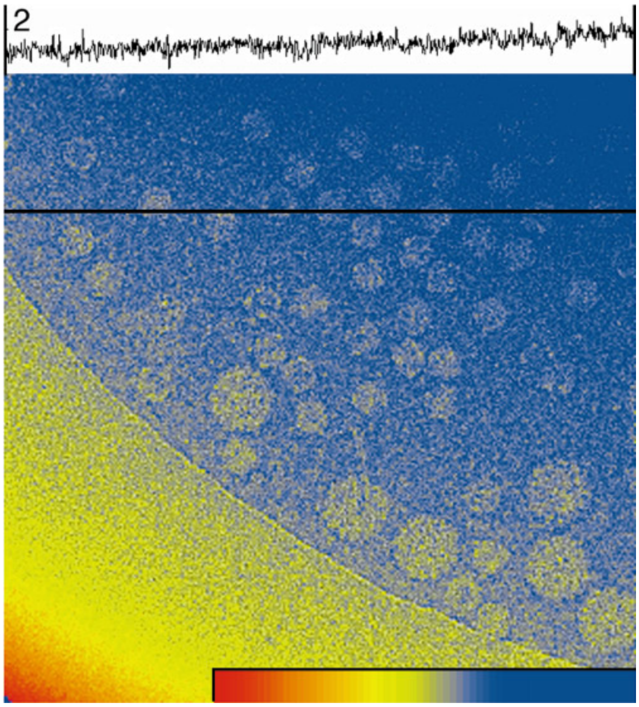


FIG. 2. Variation of the background intensity values across a digitized micrograph containing a mixture of bacteriophage ΦX174 (~30 nm in diameter) and polyoma virus (~50 nm in diameter) particles. Inset at lower right shows the color look-up table used to map image intensities. Graph at top depicts the intensity variation along a line that crosses the field of particles (black horizontal line). The virus particles are embedded in a thin (<100 nm) layer of vitrified water which is suspended across holes in a carbon film (edge seen in the lower left corner). In this example, the intensity range varies linearly from red to yellow to blue, corresponding to progressively lower densities in the specimen.

The quality of the solution can be measured in terms of the number of particles correctly identified, the number of unidentified particles, and the number of false hits. Missed particles do not constitute a severe error as long as their number is small. Information that could be gathered from these projections is lost, but the loss can be compensated by increasing the number of micrographs from which particles are selected. False hits pose a more serious type of error. If used, such regions that do not correspond to any real particle projections introduce additional errors into the 3D Fourier transform. However, methods do exist that allow such "bad" data to be screened (e.g., (*2*,*9*)).

2 AUTOMATIC PARTICLE SELECTION METHODS

Image processing of noise-obscured micrographs enables one to deal with problems such as (a) locating and extracting the motif representing the projection of a particle from a noisy background (*8*,*11*,*16*), (b) enhancing the structurally significant details of this motif (*6*,*11*), and (c) determining the orientation of the particle which produced the motif relative to some viewing direction. A number of techniques that take advantage of the high symmetry of icosahedral viruses have been successfully implemented and are routinely used to solve problems (b) and (c) (*6*,*9*). However, the human visual system remains unsurpassed in its ability to analyze micrograph images effectively and reliably.

In this section we review the effects of several image processing algorithms and heuristics on micrographs and discuss the results obtained.

2.1 Edge Detection

Edge detection is a popular segmentation method which exploits the discontinuity of the gray-level values in an image. An *edge* is the boundary between two regions with relatively distinct gray-level properties. The idea is to transform the image so that a pixel no longer contains gray-level information, but a magnitude and direction representing the severity and orientation of the local gray-level change. Gradient operators have been widely used for edge detection (*10*). A local derivative (gradient) is computed at every pixel in the image. Regions of constant intensity yield a null gradient, whereas varying regions are characterized by nonzero derivatives. The gradient of an image I at pixel (x, y) is a vector:

$$\nabla I = \begin{bmatrix} I_x \\ I_y \end{bmatrix} = \begin{bmatrix} \dfrac{\partial I}{\partial x} \\ \dfrac{\partial I}{\partial y} \end{bmatrix}.$$

A special case of gradient operators is the Sobel operators, which have both a differencing and a smoothing effect (*10*). A common implementation of the Sobel operators is (using the notations in Fig. 3a): $I_x = (z_7 + 2z_8 + z_9) - (z_1 + 2z_2 + z_3)$ and $I_y = (z_3 + 2z_6 + z_9) - (z_1 + 2z_4 + z_7)$. The masks for these operators are shown in Figs. 3b and 3c.

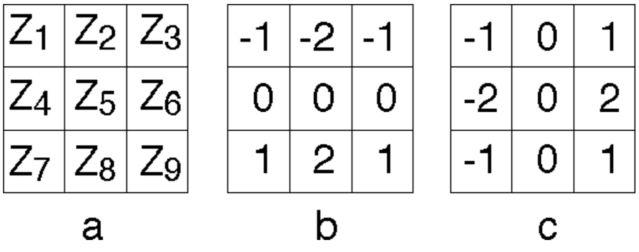
The Sobel transform fails to yield reliable results when applied to electron micrograph images. The digitized micrograph is very noisy and has significant levels of intensity variation both inside and outside particle regions. Clearly the values of the intensities alone are incapable of defining which portions of the line lie inside particles and which of them lie in the background (Fig. 2).

2.2 Template Matching

Template-matching methods have been proposed by several groups (*8*,*12*,*16*,*20*). In this method a reference (template) particle is selected from the micrograph and cross-correlated with the entire image (Fig. 4). Computationally it is more efficient to transform the entire image and the reference particle to the Fourier domain, multiply the transforms, and transform back than to perform a correlation of the original images. Peaks in the correlation pattern (i.e., values of the correlation coefficient larger than a particular threshold) identify the locations of regions in the micrograph most similar to the template and, in the ideal case, correspond to the centers of the particle projections.

Olson and Baker (*16*) proposed a two-cycle template-matching algorithm in which the peaks detected after the first correlation cycle are sorted according to their magnitude, the particle projections corresponding to the strongest peaks are averaged, and the average is used as a new template in the second cycle of the algorithm.

Template-matching methods produce reasonable results only when applied to images with a good signal-to-noise ratio, i.e., formed with medium to high electron dose, and after background variations are minimized or removed. However, it is commonly agreed that it is difficult to identify peaks in the cross-correlation maps computed from such low-dose



z_1	z_2	z_3
z_4	z_5	z_6
z_7	z_8	z_9

a

-1	-2	-1
0	0	0
1	2	1

b

-1	0	1
-2	0	2
-1	0	1

c

FIG. 3. The Sobel operator masks.

micrographs, and peak discrimination is extremely sensitive to fluctuations of the average intensity value throughout the image.

The basic template-matching algorithm described by Thuman-Commike and Chiu (*20,21*) is preceded by a constant area detection and correction process. The algorithm is rather complicated, involving image "cutting" and "sewing." The authors report percentages of false hits between 35 and 55%.

The computations involved in template-matching methods are relatively large because the complexity of the Fourier transform alone is $n \times \log(n)$, with n the number of pixels in the image. A small image may consist of $n = 1000 \times 1000$ pixels, whereas a scan of an entire micrograph can easily approach $n = 10\,000 \times 10\,000$ pixels.

2.3 The Crosspoint Method

The Crosspoint method we have developed combines traditional image processing techniques with heuristics and a new algorithm for the detection of particle centers. The time complexity of various algorithms used by this method is n, where n represents the number of pixels in the digitized micrograph. This method is described in detail in (*13*). The main steps are summarized below and illustrated in Fig. 5.

2.3.1 Image enhancement. The digitized micrograph is enhanced by histogram equalization (*10*), followed by image averaging to smooth out local fluctuations of pixel intensities. High-resolution 3D reconstructions usually include close-to-focus, i.e., low-contrast, images in which the high-resolution details are not destroyed by the electron beam. Histogram equalization helps improve image contrast by redistributing the gray levels in the image more uniformly over the gray-scale range (Figs. 5a–5c).

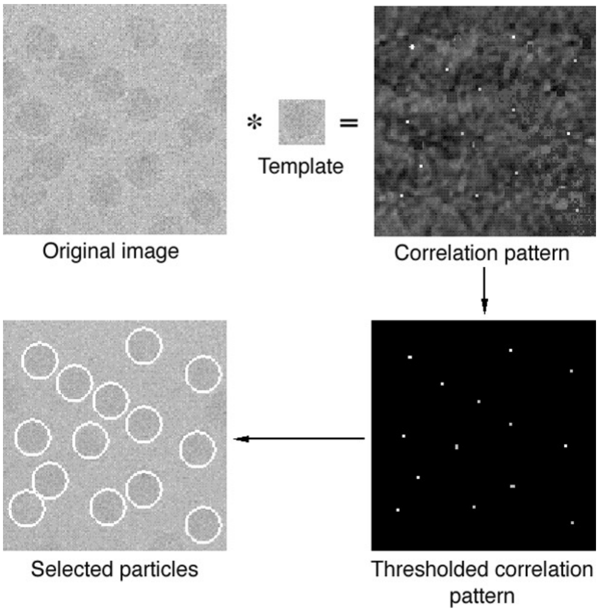


FIG. 4. The basic cross-correlation, template-matching algorithm.

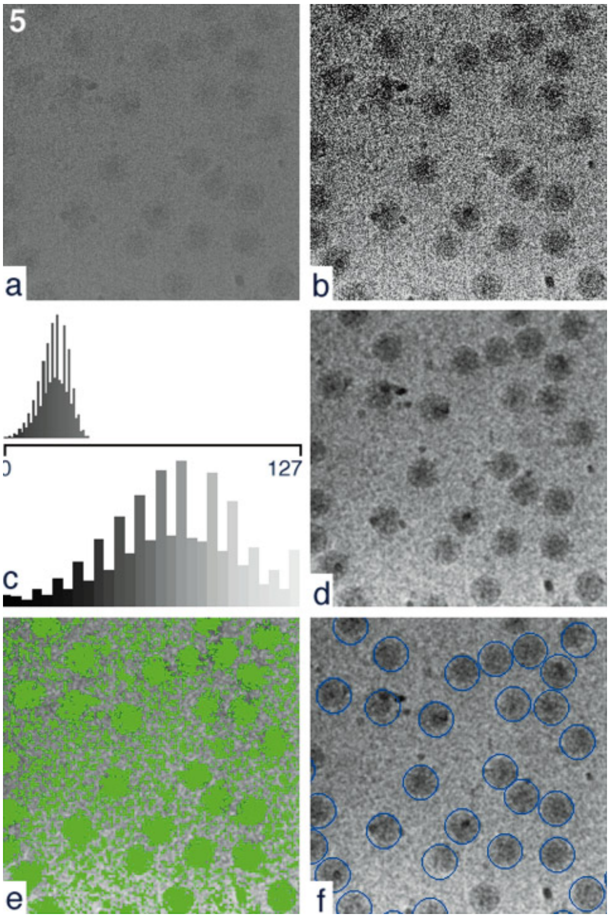
The rationale for neighborhood averaging the digitized, histogram-equalized image is motivated by the fact that the intensities of the pixels in the image are not characteristic of the inside or the outside of a particle projection (see Fig. 2). A particular intensity value may occur in a region inside a projection as well as somewhere in the background, where there are no particles. However, the majority of the pixels inside a particle projection have lower intensity (are darker) than the pixels surrounding the particle, thus enabling the human eye to easily recognize particles. High-intensity fluctuations tend to be sharp and scattered throughout the entire area of the projected image. Since averaging is a smoothing operation, such fluctuations are reduced or disappear completely in this process. However, the size of the averaging filter must be chosen carefully to prevent the resulting image from becoming too blurred (Fig. 5d).

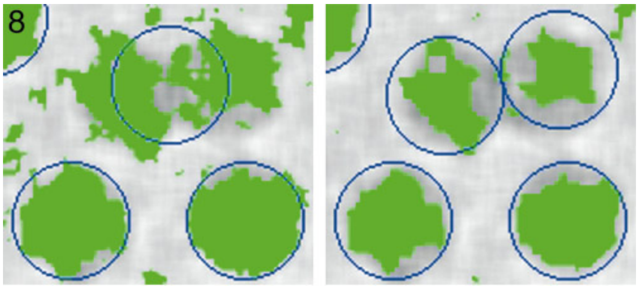
2.3.2 Particle identification with a double scan procedure. The particle identification algorithm is at the core of the Crosspoint (CP) method. It consists of two phases: *marking* and *clustering.*

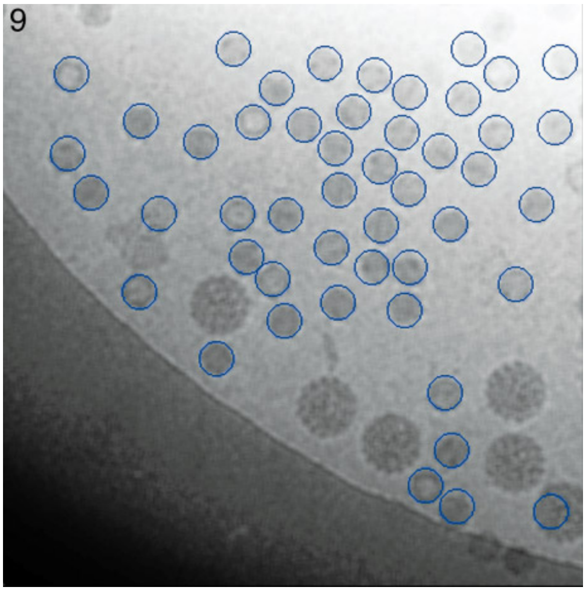
The algorithm takes as input an image and the value of the radius r of the particles to be identified. The image could be the original raw image, the enhanced image, or a subimage. The radius of the particles is defined interactively based on the visual inspection of the image or is inferred from measurements made on previous images. The result of the marking phase is a binary image. Each pixel is considered to belong either to a particle projection (marked, set to 1) or to the background (not marked, set to 0).

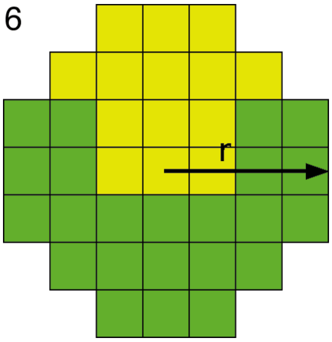
In the original CP method (*13*) the image is scanned horizontally, row by row from top to bottom. Pairs of pixels at distance $r + 1$ are compared and the difference between the intensity values of the pixels in such a pair is tested against a threshold value. If this difference is larger than the threshold, the algorithm proceeds to compare the lower intensity value with that of a pixel at distance $r + 1$ in the vertical direction. This difference is also tested against the threshold and, if it is larger, the algorithm marks the element of the pair with the lower intensity as being inside a particle; otherwise the pixel is not marked.

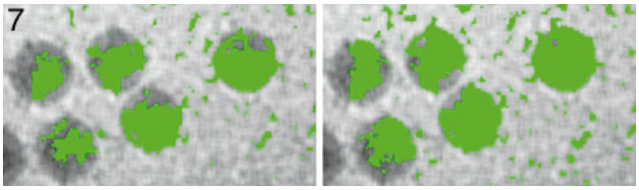
A portion of a micrograph after marking (Fig. 5e) shows pixels that have been marked as being inside a particle colored in green, and those unmarked











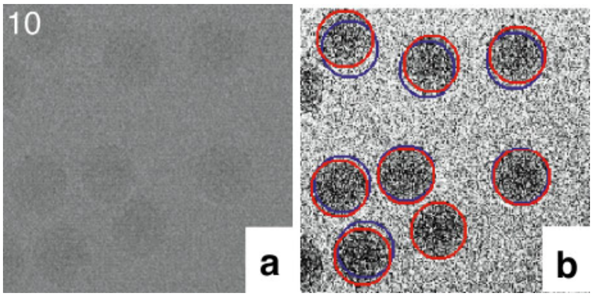


FIG. 5. (a) Portion of a low-contrast micrograph of frozen-hydrated sample of reovirus cores. (b) The micrograph after histogram equalization. (c) Gray-level histograms before (top) and after (bottom) histogram equalization. (d) The micrograph in (b) after neighborhood averaging with a 10 × 10 filter. (e) Contents of the binary image after pixel marking (green) superimposed on the micrograph in (d). (f) The result of the CP2 method.

FIG. 6. The result of the marking phase in the case of an ideal particle. *Note.* The three lowest yellow pixels should have been colored green.

FIG. 7. Portion of a micrograph in which the pixels have been marked (a) once and (b) twice. In (b), the particle projections, and hence their centers, are better approximated by the clusters.

FIG. 8. Disconnecting particles by thinning: (a) particle identification without thinning and (b) particle identification with thinning.

FIG. 9. The solution produced by the CP2 method for the micrograph shown in Fig. 2.

FIG. 10. (a) Electron micrograph containing several particle projections. (b) New particle positions (in red) after refinement (old positions—in blue—are shown for comparison).

retaining their original intensities. For most particles, the clusters of green pixels approximate quite well the area of the particle's projection. The center of each particle is computed as the center of mass of the cluster corresponding to that particle.

Owing to the asymmetric nature of the scanning process, the top region of each particle projection is systematically left unmarked. The pixels marked by the algorithm in the case of an ideal particle (Fig. 6) are colored green, whereas the top portion of the particle (yellow) is not marked because the comparison between the intensities of those pixels and the ones at distance $r + 1$ in the vertical direction fails. Hence, this single-scan process results in a systematic misjudging of the particle centers in the vertical direction.

A more accurate version of the algorithm, CP2, involves scanning the image twice: the first scan is performed as before, followed by a second scan applied to a transposed image rowwise (i.e., from bottom to top). The marked pixels are the cumulative sum of both scans. A portion of a micrograph with pixels marked (a) after CP and (b) after CP2 is illustrated in Fig. 7.

Clustering is the second phase of the particle identification algorithm. It determines the clusters, i.e., the connected components in the binary image resulting from the marking phase. Two algorithms, one based on a depth-first search and the other on a coloring scheme, are briefly described.

The *stack algorithm* (*13*) is a depth-first algorithm for detecting connected components in a binary array using a stack. The array is scanned rowwise until the first 1 is encountered. Its coordinates are used to update the center of mass of the cluster currently being determined and the size of the cluster is incremented. All marked neighbors of the current position are pushed onto a stack (eight neighbors are considered). The next position to be processed is the one at the top of the stack. A cluster has been completely detected and processed when the stack becomes empty. The horizontal scanning of the binary array then resumes, until all clusters have been detected.

The *coloring algorithm* was suggested by M. J. Atallah (*1*). It detects connected components in a binary array by "coloring" them with different colors. As in the previous algorithm, the array is scanned rowwise, and every time a marked position is encountered, it is either colored with a new color from a color array (if none of its neighbors is colored) or it receives the color of its neighbors. Only four neighbors are considered (top left, top, top right, and left). A decision must be made when, at some point during the scanning process, two clusters that have been considered separate and have been colored with two different colors become connected. In this case, the two clusters have to be "recolored" with the same color. The simplest way to achieve this is to make the two different colors synonyms. The center of mass and the size of the clusters can be computed on-the-fly, as the scanning progresses.

The size of a cluster is used to filter out clusters that are too large or too small compared with the expected area of the projection (see Section 3.2 and (*13*)). The center of mass of each cluster approximates the corresponding particle center. The application of CP2 to the micrograph in Fig. 5e is shown in Fig. 5f.

2.3.3 Postprocessing. A particle identification method is affected by two types of errors: (a) missed particles and (b) false hits. In the CP2 procedure the number of missed particles can be reduced by adjusting the rejection criteria based on the size of the clusters. A false hit occurs when a cluster that does not correspond to a real particle projection is accepted. As mentioned in Section 1, this is a more serious type of error. One way to reduce the number of false hits is to calculate the average intensity inside each of the particles detected and to compare it with the average of all intensity values in a circular region outside it. If the two average values are very close, then it is very likely that the particle is merely a false hit.

One of the most common causes for missing particles in the CP2 method occurs when just one cluster is detected instead of two, as results when two particle projections are touching or are very close to one another (see Fig. 8a). Here, postprocessing is necessary to disconnect the two clusters. A "thinning" procedure has been adopted in which the outermost layers of pixels from each cluster are removed and this can effectively disconnect the clusters that have merged into single clusters. This procedure works in such situations because the clusters are generally connected by thin "bridges." The appearance of clusters before and after the thinning procedure is illustrated in Fig. 8.

Results produced by the CP2 method for a micrograph which contains a mixture of two types of virus particles is shown in Fig. 9. In this example, the desired particles were the smaller ones (bacteriophage ΦX174). The larger particles (polyoma virus) were included solely for calibration purposes (*16*). The image is particularly noisy, a large portion of the carbon film obscures the lower left corner, and the variation of the background intensity is clearly visible. Nevertheless, the CP2 method is quite successful in identifying most of the ΦX174 images and distinguishes them from other objects (polyoma particles, carbon film, unidentified contaminants).

3 THE REFINEMENT OF PARTICLE POSITIONS

Our experience with a large number of micrographs of different virus samples indicates that the centers determined using the CP2 method approximate fairly well the true centers of the particles. However, it is possible to improve the quality of our solution by refining the centers determined by CP2. We assessed the quality of our refinement method by comparing the positions of the centers before and after refinement with centers selected by an experimentalist. In the case of Fig. 10a, two corresponding positions differ, on average, by (1.65, 3.37) pixels. The error function used to calculate the average distance between a position (x_i^a, y_i^a) detected by the program and its manually selected counterpart (x_i^m, y_i^m) is given by the formula

$$(ex, ey) = \left(\sqrt{\sum_{i=1,\ldots,C} (x_i^a - x_i^m)^2} \Big/ C, \; \sqrt{\sum_{i=1,\ldots,C} (y_i^a - y_i^m)^2} \Big/ C \right),$$

where C is the total number of particle images present in the micrograph.

The next section describes a correlation-based method for refining the centers of the particles obtained using the CP2 method and also analyzes the results obtained. A second, background equalization method was tested but produced unreliable results.

3.1 Correlation-Based Refinement of Particle Centers

This algorithm was inspired by the template-matching algorithms (*16,20,21*), but it is more efficient and accurate. Let C be the number of particle projections detected using the CP2 method and let r be the radius of the particles. Provided the number of false hits and of missed particles is small, a model particle projection built by averaging all the C projections is more accurate than one constructed manually by selecting one or few particles as described in (*16*). The model can be cross-correlated with the points of the entire scanned image, but restricted to a limited search region. We take advantage of the accuracy of the CP2 solution by restricting this region to a small area around the center of each of the C particles. For a square-shaped region of dimension $2b + 1$ pixels where b is typically set to a value in the range of 2 to 4, the total number of correlations performed is $C(2b + 1)^2$. The position yielding the largest correlation coefficient identifies the region in the micrograph most similar to the reference and it is likely to be a more accurate approximation of the true center. The two steps of the procedure are

Step 1. Build the model particle projection. The intensity of every pixel of the model is the average of the C corresponding pixel intensities in all detected particle images.

Step 2. For each of the C particles, for every position (x_b, y_b) within the search box, correlate the model particle with the micrograph in a circular region of radius r centered at (x_b, y_b).

Let I denote the image on which the refinement is to be performed, M the model particle, $\bar{I}_b$ and $\sigma_{I,b}$, respectively, the average intensity and the standard deviation of I inside a circle of radius r centered at (x_b, y_b), N the number of pixels inside the model particle, $\overline{M}$ the average intensity, and σ_{M} the standard deviation of the model particle. The correlation coefficient, ρ, is given by the formula (*15*)

$$\rho = \frac{\sum_{(x_i,y_i)\in M} (I(x_i + x_b, y_i + y_b) - \bar{I}_b) \times (M(x_i, y_i) - \overline{M})}{N \times \sigma_{I,b} \times \sigma_M}.$$

The new, refined center corresponds to the position yielding the largest ρ. We have implemented this algorithm such that if the new center is located on the border of the search box, we allow for the search box to move in the direction of the maximum correlation coefficient a number of times to obtain a more accurate value for ρ.

The results of the correlation refinement for a test image are shown in Fig. 10b. The particle center coordinates after refinement approximate the true centers better than the initial values. Analysis of a large number of micrographs shows that correlation with a model particle usually improves the results produced by the Crosspoint method. For the image shown in Fig. 10a, the average error is (0.83, 0.84) pixels for the correlation-based refinement, as opposed to the (1.65, 3.37) pixel error obtained before refinement. The time to produce the initial CP2 solution in the case of a 1280 × 1000 pixel image (Fig. 12a) was about 18 sec on a Reality Engine Silicon Graphics machine with a 90-MHz processor and 128 Mbytes of main memory. The subsequent correlation refinement step took about 15 sec.

TABLE 1

Sensitivity of the Crosspoint Method to Changes in the Particle Radius

Radius (in pixels)	Correctly identified	Missed	False hits
18	36 (77%)	11 (23%)	3 (6%)
20	39 (83%)	8 (17%)	2 (4%)
22	38 (81%)	9 (19%)	2 (4%)
24	42 (89%)	5 (11%)	2 (4%)
25	42 (89%)	5 (11%)	2 (4%)
26	39 (83%)	8 (17%)	3 (6%)
28	40 (85%)	7 (15%)	3 (6%)
30	39 (83%)	8 (17%)	6 (12%)

Note. Results are given for the micrograph shown in Fig. 9. The true radius is approximately 25 pixels.

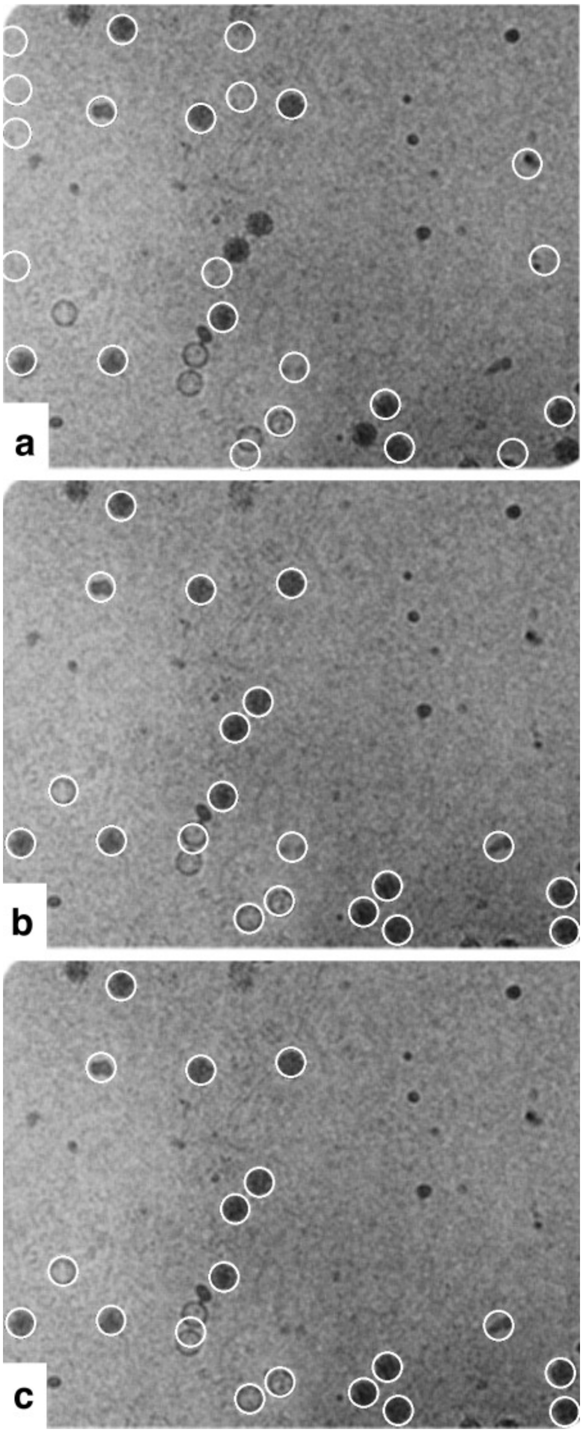


FIG. 11. Sensitivity of the Crosspoint method to changes in the number of thinning layers: (a) no thinning, (b) one thinning layer, (c) two thinning layers. The micrograph shown contains images of human rhinovirus (HRV) particles decorated with Fab antibody fragments.

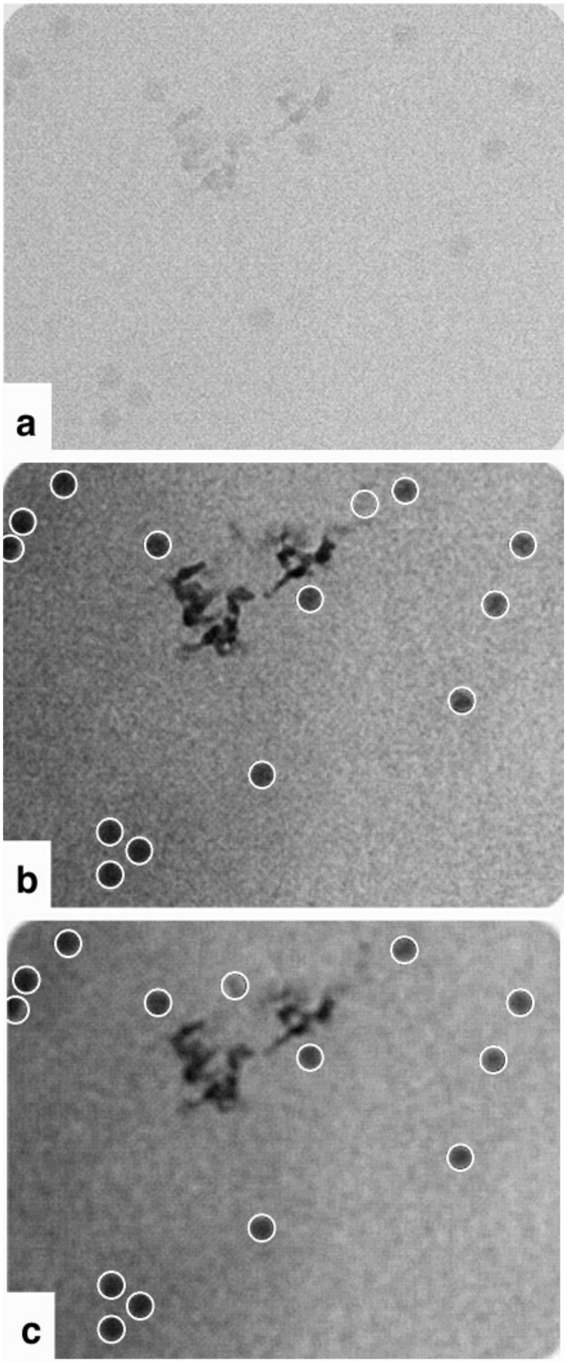


FIG. 12. Sensitivity of the Crosspoint method to changes in the image pixel resolution: (a) original low-contrast image (1280 × 1000 pixels) showing projections of several human rhinovirus particles, (b) the result of the CP2 method applied to (a), (c) the result of the CP2 method applied to (a) after subsampling (640 × 500 pixels).

We also designed and tested a background equalization refinement method, but with unsatisfactory results on our test images. In this method, the solution generated by the CP2 method was used to produce an initial set of particle centers, at which point background variations were removed before correlation-based refinement of the centers was performed.

3.2 The Sensitivity of the Crosspoint Method

The CP2 method is sensitive to changes in several parameters. For example, the radius *r* of the virus particle projections is a very important input parameter. The CP2 method cannot be used for micrographs containing a mixture of different virus particles that are comparable in size (e.g., whose diameters differ by only ∼10%). Often it is difficult even to locate the particles in the original image; hence, errors in defining the radius are expected.

Our experience indicates that the quality of the solution is not seriously affected for small variations of *r*. Table 1 illustrates the results obtained for one micrograph (Fig. 9).

Another solution-sensitive parameter, set by the program user, is the number of thinning layers used to disconnect particles that have joined into a single cluster. An illustration of the use of the Crosspoint method using zero, one, and two thinning layers, respectively, is shown in Fig. 11. For most micrographs we have tested, two thinning layers are optimal.

Three other parameters influence the CP2 solution. One is the threshold used in the marking phase of the particle identification algorithm. The other two are the upper and lower bounds for the size of a cluster of marked pixels. A cluster is considered to represent a particle if its size approximates the area of a circle with the same radius as the particle. We have selected optimal values for these bounds based upon our analysis of a large number of micrographs.

4 PROCESSING LARGE IMAGES

Modern transmission electron microscopy methods now make it possible to produce very large images, with 50–100 Mpixels (million pixels). One micrograph may contain the projected images of a thousand or more virus particles. Manipulating such an image in the computer requires 50–400 Mbytes of storage depending on the dynamic range of the imaging device (1, 2, 3, or 4 bytes/pixel). The fact that such images can be generated at a fairly high rate and have to be stored creates a critical need for large secondary storage systems and data compression techniques.

Processing and rendering such images is a challenging proposition due to the speed and storage limitations of current graphics workstations. For example, rendering a 5878 × 7521 pixel image takes about 120 sec on an SGI Power Onyx with one processor and 128 Mbytes of memory. Histogram equalization of the same image takes more than 200 sec, and 10 × 10 averages more than 400 sec.

Several possible solutions to this problem exist. A graphics system with several processors and 512 Mbytes to 1 Gbyte of main memory could be used. Parallel algorithms for image enhancement, particle identification, and center refinement are needed to exploit efficiently such an expensive machine. Alternatively, the large image can be cut into subimages and each subimage then processed independently.

Another option is to compress the original, digitized image and then to detect the initial positions of the particles on the compressed image. Once these positions are detected, one can cut the original image more efficiently and conduct the refinement procedure on each of the subimages, using as an initial approximation the positions located on the compressed image.

4.1 Image Compression

The size of an image can be significantly reduced by simple compression algorithms. A lossless compression method like run-length encoding (RLE) is not very useful because it alters the contents of the image and the Crosspoint method is not designed to work on an encoded image.

The alternative to lossless compression is lossy compression. Several lossy compression schemes are possible. The simplest one is subsampling. Image size can be decreased by a factor of m^2 by using only every mth pixel on each row and every mth row of the image. More sophisticated algorithms involving

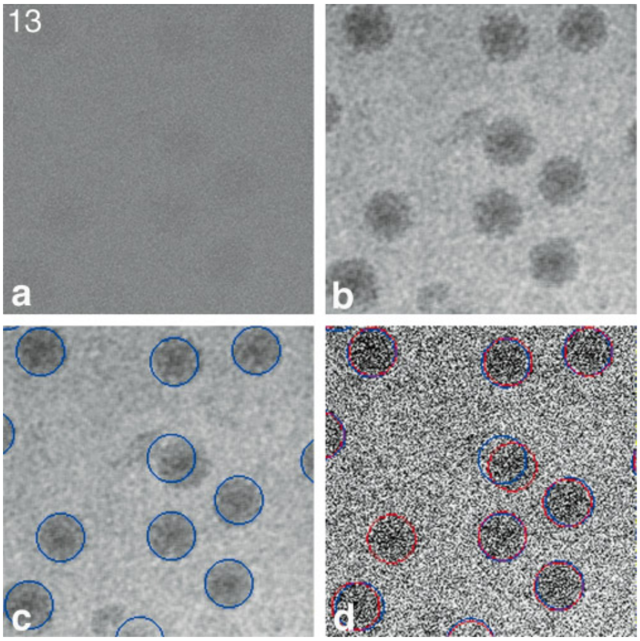


FIG. 13. Experimenting with a low-contrast image: (a) portion of a 5878 × 7521 pixel image; (b) image after histogram equalization and averaging; (c) CP method applied to image in (b); (d) refinement of the positions detected in (c): blue circles—positions before refinement; red circles—positions after refinement.

wavelet transformations (*19*) could also be used to compress micrograph images. Preliminary results indicate that such transformations can be successfully used even for very low-contrast images in conjunction with the Crosspoint method (Fig. 12).

4.2 A Particle Identification Algorithm Based on Image Subsampling and Decomposition

In contrast with correlation-based methods for which higher pixel resolution means better chances for a more accurate match with the template, the Crosspoint algorithm works well on subsampled images. A 1280 × 1000 pixel micrograph containing several human rhinovirus particle images (Fig. 12a) was examined with the CP2 method on the original image (Fig. 12b) and on the subsampled image (Fig. 12c; subsampling factor, $m = 2$). The average error for the coordinates of the particle centers selected manually versus using the CP2 method in the case of Figs. 12b and 12c, without refinement, is (0.34, 1.04) pixels for the full-resolution image and (0.49, 1.25) pixels for the subsampled one.

The processing of a large image involves the following steps:

Step 1. Reduce the size of the image by a factor of m^2. Values of $m = 2$ and $m = 3$ (i.e., four- and ninefold subsampling) seem sufficient for all practical purposes.

Step 2. Enhance this image by histogram equalization and averaging.

Step 3. Specify the radius r of the particles to be identified and use the Crosspoint method to identify them on the image resulting after Step 2. Let (x_i, y_i), $i = 1, \ldots, C$, denote the coordinates of the C particle centers detected.

Step 4. Divide the original image into P subimages. Several strategies can be used. One is to ensure that each subimage has a rectangular shape and contains roughly C/P particles. Another is to divide the image into subimages of equal size (possibly overlapping).

Step 5. For each subimage carry out the correlation-based refinement algorithm described in Section 3.1. Construct a model particle projection by averaging the particles within that subimage. Allow the center of particle i to move within the box with corners $(x_i - b, \mathrm{y}_i - b)$, $(x_i - b, y_i + b)$, $(x_i + b, y_i - b)$, and $(x_i + b, y_i + b)$. As before, if the best correlation is obtained when the center is located on the edge of the refinement box, allow up to k moves of the refinement box.

Step 6. Filter out particles whose best correlation coefficient is lower than a given threshold (likely to be false hits).

The timing results for the CP2 method in the case of the micrograph in Fig. 12 recorded on a Silicon Graphics workstation with a 200-MHz IP22 processor and 64 Mbytes of main memory were 18 sec for the entire image (1000 × 1280 pixels) and 7 sec for the subsampled image (subsampling factor $m = 2$).

4.3 Image Decomposition

An alternative to image compression is to decompose it into several overlapping subimages and to apply the Crosspoint method to each subimage independently. With the exception of histogram equalization, the CP2 procedure does not involve any global image transformation (such as Fourier transforms) and it is, therefore, suitable to be applied to individual subimages. Histogram equalization is the only transformation affected by decomposition. By applying histogram equalization to subimages an improved contrast is obtained in each subimage, closer to the optimal image contrast that would be achieved by using, for instance, the adaptive histogram equalization technique described in (*17*).

In contrast to the complex "cutting" and "sewing" described in (*20,21*), the only requirement of the method proposed here is that each subimage include a border region to allow for correct averaging, clustering, and box migration during the refinement stage for all particles inside. Let r denote the radius of the particles, $2 \times b + 1$ the length of the side of the refinement box, and k the maximum number of box shifts allowed during the refinement. Then, the width of the border region should be $w = 2 \times r + max\{r + 1, k \times b\}$. For example, a 10 000 × 10 000 pixel image, with $r = 64$, $b = 4$, and $k = 5$, can be decomposed into four subimages of 5193 × 5193 pixels each. In this case, the border region has the width $w = 2 \times 64 + max\{65, 5 \times 4\} = 193$ pixels and the actual subimage has 5000 × 5000 pixels. Due to the need to include a border region, there is a point of diminishing return when increasing the number of subimages into which an image is decomposed.

Special attention must be paid to processing clusters located close to the border region. If a cluster is fully contained within the extended subimage and its center of mass is within the boundaries of the actual subimage then it is considered to belong to the subimage. If a cluster is fully contained in the extended subimage, but its center of mass is inside the border region, the cluster is not considered part of the current subimage. Its processing will take place in one of the neighboring subimages.

Due to the nature of the decomposition and the fact that the center of a cluster may belong to only one of the actual subimages, each cluster is processed only once. Therefore, it is straightforward to combine the results from all subimages: the list of all particle positions for the whole image is the union of all subimage lists.

5. AN ENVIRONMENT FOR EXPERIMENTING WITH PARTICLE IDENTIFICATION METHODS

The expectation that one can design a fully automated particle identification method capable of processing micrographs produced under various conditions without any human intervention seems unrealistic at this time. What we believe can be done at this stage is to design an environment which supports experimenting and tuning of various methods.

Given a batch of images obtained under similar conditions, the user needs to fine tune the general algorithm, e.g., the number of thinning layers, the radius of the particle, etc. Once an optimal procedure is established, all images in the batch can be processed automatically. Occasionally, a different sequence of image enhancement steps leads to better results than the one we described in Section 2.3.1. Figure 13a shows an image in which particles can be identified with the naked eye only by a very astute observer. On the enhanced images (Figs. 13b and 13c), the particles can be identified and the CP method works well. In this case, the following sequence of image enhancement steps leads to the best results: histogram equalization followed by averaging, followed by another cycle of histogram equalization and averaging steps.

To support such experimentation, the environment we have developed supports the standard particle identification method described earlier, as well as individual image transformations that can be composed in random order.

EMMA is an interactive software package built around the Crosspoint method. In addition to automatic particle selection and refinement, it includes capabilities to decompose large images and to display the subimages, to perform various traditional image processing transforms on the digitized micrographs, to select, unselect, and extract individual particles interactively, and to store particles into files. The transforms supported are: histogram equalization, averaging, Sobel and Laplace gradient methods, high-boost filtering, colormap modification, compression, and the Hough transform. The program allows for an easy composition of such transforms in the order specified by the user.

EMMA is built in X-Windows and Motif (*21*) and consists of approximately 20 000 lines of code.

6 CONCLUSIONS AND FUTURE WORK

To improve the resolution of virus structures determined using cryoelectron microscopy methods from 20 Å to 5–10 Å, the number of virus particle projections used in the three-dimensional reconstruction process must increase from a few hundred to several thousands. To make better use of the biological samples, the electron microscope must be controlled to aim its beam at particles with positions previously determined from low-dose, low-magnification, and hence very noisy images. Modern devices are capable of producing images with 100 Mpixels, or even more, containing thousands of virus particle projections that need to be analyzed, hence, the motivation for the work reported in this paper.

Efforts to automate the particle identification process have been reported in the literature, but existing methods are inefficient and none of them has gained wide acceptance. Noise due to a variety of sources makes particle identification very difficult. Due to the low contrast of some of the micrographs, it is often a challenge for the human eye to even notice a particle in a micrograph. And it is not an easy task to capture in an algorithm the eye's ability to recognize shapes.

The original automatic particle identification method, the Crosspoint or CP, is described in (*13*). As reported in (*13*), the algorithm is efficient and produces relatively accurate solutions, with acceptable miss rates and few false hits. By exploiting the local properties of the micrograph image, the algorithm is capable of dealing with a varying background.

After a large number of tests, we conclude that the particle identification algorithm works best on images enhanced by histogram equalization and averaging. In this paper, we propose a refinement method based on correlating a model particle with regions of the image located in the vicinity of the particles detected by the CP algorithm.

Another contribution of this paper is an algorithm for processing large images. A compressed image is used to obtain the initial list of particle centers. Experiments confirm that a 4- to 16-fold lossy compression does not deter the CP algorithm from locating particle centers with sufficient accuracy. The uncompressed large image is then decomposed into subimages to which the refinement algorithm is applied.

There are no obvious ways to compute the quality of the solution provided by a particle identification program. The best one can do is to identify manually the location of particles on one micrograph, compare them with those computed by the program, and report the error. It is difficult to compare different algorithms and programs. There are no benchmark images and it is possible that a program which does very well on some images may provide a poor-quality solution for others. Likewise, there are few timing results to allow a fair comparison of different programs. Nonetheless, an analysis of the algorithms involved favors the CP method over methods requiring Fourier transforms.

The speed of such a method is a prerequisite for automatic control of the electron microscope. To avoid premature damaging of the biological specimen, the following approach can be used: first, a low-dose, low-magnification image is recorded on a slow-scan CCD camera and the digital image is used to determine the coordinates of the virus particles; then these coordinates are used to aim and calibrate the electron beam to take high-magnification pictures of each particle or clusters of particles, using flood-beam or spot-beam imaging procedures (*7*).

Further information about the EMMA package, as well as a number of test images, can be obtained from http://www.cs.purdue.edu/homes/sb/Projects/EMMA/emma.html. The software is available free upon request.

This research has been partially supported by the National Science Foundation Grants BIR-9301210 and MCB-9527131, by a grant from the Intel Corporation, a grant from the Purdue Research Foundation, a Grant-In-Aid of Research and a Summer Faculty Fellowship from Indiana University, and by the Scalable I/O Initiative. We thank the anonymous reviewers for many constructive comments.

REFERENCES

1. Atallah, M. J. (1996) Private communications.

2. Baker, T. S., and Cheng, R. H. (1996) A model-based approach for determining orientations of biological macromolecules imaged by cryo-electron microscopy, *J. Struct. Biol.* **116,** 120–130.

3. Böttcher, B., Wynne, S. A., and Crowther, R. A. (1997) Determination of the fold of the core protein of hepatitis B virus by electron microscopy, *Nature* **386,** 88–91.

4. Cheng, R. H., Kuhn, R. J., Olson, N. H., Rossmann, M. G., Choi, H.-K., Smith, T. J., and Baker, T. S. (1995) Three-dimensional structure of an enveloped alphavirus with $T = 4$ icosahedral symmetry, *Cell* **80,** 621–630.

5. Conway, J. F., Cheng, N., Zlotnick, A., Wingfield, P. T., Stahl, S. J., and Steven, A. C. (1997) Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy, *Nature* **386,** 91–94.

6. Crowther, R. A., DeRosier, D. J., and Klug, A. (1970) The reconstruction of a three-dimensional structure from projections and its applications to electron microscopy, *Proc. R. Soc. London Ser. A* **317,** 319–340.

7. Downing, K. H. (1991) Spot-scan imaging in transmission electron microscopy, *Science* **251,** 53–59.

8. Frank, J., and Wagenknecht, T. (1984) Automatic selection of molecular images from electron microscopy, *Ultramicroscopy* **12,** 169–175.

9. Fuller, S. D., Butcher, S. J., Cheng, R. H., and Baker, T. S. (1996) Three-dimensional reconstruction of icosahedral particles—The uncommon line, *J. Struct. Biol.* **116,** 48–55.

10. Gonzalez, R. C., and Woods, R. E. (1993) Digital Image Processing, Addison–Wesley, Reading, MA.

11. Harauz, G., and Fong-Lochovsky, A. (1989) Automatic selection of macromolecules from electron micrographs by component labeling and symbolic processing, *Ultramicroscopy* **31,** 333–344.

12. van Heel, M. (1982) Detection of objects in quantum-noise-limited images, *Ultramicroscopy* **8,** 331–342.

13. Boier Martin, I. M. (1996) Scientific Data Visualization and Digital Image Processing for Structural Biology, Ph.D. thesis, Purdue University.

14. Moody, M. F. (1990) Image analysis of electron micrographs, *in* Biophysical Electron Microscopy, pp. 145–285, Academic Press, New York.

15. Mosteller, F., Rourke, R. E. K., and Thomas, J. B., Jr. (1970) Probability With Statistical Applications, Addison–Wesley, Reading, MA.

16. Olson, N. H., and Baker, T. S. (1989) Magnification calibration and the determination of spherical virus diameters using cryo-microscopy, *Ultramicroscopy* **30,** 281–298.

17. Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. (1987) Adaptive histogram equalization and its variations, *Comp. Vis. Graph. Image Proc.* **39,** 269–280.

18. Smith, P. R., Peters, T. M., and Bates, R. H. T. (1973) Image reconstruction from finite numbers of projections, *J. Phys.* **6,** 319–381.

19. Stollnitz, E. J., DeRose, A. D., and Salesin, D. H. (1996) Wavelets for Computer Graphics: Theory and Applications, Morgan Kaufmann, San Mateo, CA.

20. Thuman-Commike, P., and Chiu, W. (1995) Automatic detection of spherical particles from spot-scan electron microscopy images, *J. Microsc. Soc. Am.* **1,** 191–201.

21. Thuman-Commike, P. A., and Chin, W. (1996) PTOOL:A software package for the selection of particles from electron cryomicroscopy spot-scan images, *J. Struct. Biol.* **116,** 41–47.

22. Young, D. A. (1990) The X Window System Programming and Applications with Xt, the OSF/Motif ed., Prentice Hall, New York.