Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Structural Biology 145 (2004) 111–122
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

Classical detection theory and the cryo-EM particle selection problem

Fred J. Sigworth*

Department of Cellular and Molecular Physiology, Yale University, 333 Cedar Street, New Haven, CT 06520, USA

Received 6 August 2003, and in revised form 16 October 2003

Abstract

Particle selection is an essential but tedious step in the determination of macromolecular structures by single particle reconstruction. This paper presents an automatic, multi-reference particle detection scheme that is based on the classical matched filter principle. It makes use of a pre-whitening filter to standardize the noise, a reduced representation of the references by means of principal component analysis, and a statistic for distinguishing particles from image artifacts. Standardizing the noise allows the noise-induced false-positive frequency to be estimated, and also allows the distribution of the discrimination statistic to be calculated a priori. The method is demonstrated with an annotated dataset of cryo-EM images.

1. Introduction

The identification of individual particles in micrographs is a bottleneck in the high resolution determination of protein structures by electron cryomicroscopy (Nicholson and Glaeser, 2001). There are two goals that must be satisfied by a successful implementation of automatic particle picking. The first is to detect individual particle images in the presence of random noise. This is a problem that is equivalent to the classical problem of detecting symbols in a noisy communication channel. The second goal is to distinguish true particle images from artifacts or images of corrupted particles. This is a more difficult problem because in general its solution requires both information about the desired particles and also information about the characteristics of non-particles.

In this paper an implementation of automatic particle picking is based on the classical "matched filter" or "correlation detector" method. The prerequisite is that some sort of 3D density map of the particle is already available; projections of this map are used to derive references for the detector. In the past, correlation-based schemes using a single reference (Frank and Wagenknecht, 1984; Lata et al., 1995) or a representative set of references (Ludtke et al., 1999; Roseman, 2003; Stoschek and Hegerl, 1997; Wong et al., 2003) have been presented. The approach followed here uses the classical matched filter rather than the modern statistical techniques employed by Stoschek and Hegerl (1997) and by Wong et al. (2003). The main difference from previous work is that the spectrum of the background noise of a micrograph is standardized through the application of a pre-whitening filter. With this standardization, the frequency of finding "false particles" can be estimated, and a statistic for discriminating "true" particles has a known distribution. Also described here is a method to reduce the computational burden of using many references, employing principal component analysis (PCA). A closely related application of PCA is described by Ogura and Sato (2003) for their particle picker that is based on a neural network.

2. The particle detection problem

To illustrate the algorithms described here, the results will be given with reference to an annotated keyhole limpet hemocyanin (KLH) dataset (Zhu et al., 2003) available at http://ami.Scripps.edu/prtl_data. This dataset consists of 82 micrographs, each 2k × 2k pixels in size, with the pixel size being 2.2 Å. For the processing described here the high-defocus "Exposure 2" micrographs were used, and binning of pixels was employed to

* Fax: 1-203-785-4951.
E-mail address: fred.sigworth@yale.edu.

1047-8477/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2003.10.025

increase the pixel size to 8.8 Å. The images show "side" views and "top" views of the KLH particles. In this example only side views are picked.

The following notation will be used. An upper-case variable denotes an image, that is a 2D array of pixel values. Given an image A, the notation $A(\mathbf{r})$ represents the value of the pixel at the location $\mathbf{r}$ in image coordinates, while a_i is the value of the ith pixel when pixels are addressed lexicographically. The inner product of two images, each with n_p pixels, is defined as

$$X \bullet Y = \sum_{i=1}^{n_p} x_i y_i,$$

while the pixel-by-pixel product of the images, is denoted XY. The squared norm or "power" of an image is defined as

$$|X|^2 = X \bullet X.$$

2.1. Noise model

The particle detection problem is simply to locate images of particles randomly distributed in a micrograph. The classical approach to this problem is to start with a statistical description of the background noise. The simplest case is to assume that the noise is independent and identically distributed, Gaussian, and additive. Noise that is independent from pixel to pixel has a flat power spectrum. Noise having a Gaussian distribution has properties that greatly simplify the analysis. For example, linear filtration of Gaussian noise yields Gaussian noise; also, a maximum-likelihood estimator of a Gaussian random variable is the same as a least-squares estimator. The "additive" property means that the micrograph can be modeled as a mosaic of noiseless particle motifs to which the noise is added. We will exploit this collection of properties in the following analysis.

An image of a particle is modeled as the sum of the particle motif A and independent, identically distributed, zero-mean Gaussian pixel noise N

$$X = mA + N, \tag{1}$$

where m is the motif amplitude. The pixels of the motif A are assumed to be non-zero only inside the particle boundary, and are scaled such that $|A|^2 = 1$; further, at each pixel the variance of the noise N is assumed to be unity.

In an actual micrograph the main source of fluctuation is shot noise from the small number of imaging electrons (for example, 10–20 electrons per square angstrom of specimen area). This shot noise follows a Poisson distribution, but a Gaussian distribution is a good approximation when each pixel represents many square angstroms. The shot noise is intrinsically white; that is, it has a flat power spectrum, which means that the noise in individual pixels is uncorrelated. However in practice the background noise in a cryo-EM micrograph is not white, for two reasons. First, inelastically scattered electrons form an out-of-focus image that contributes to low frequencies. Variations in the thickness of ice contribute large low-frequency components, and an amorphous carbon support film also contributes low-frequency noise. Second, there is a fall-off of noise (and signal) at high frequencies due to the point-spread function of the camera or electron detector.

A simple way to compensate for these effects is to create a "pre-whitening" filter. Blank areas of micrographs are used to compute the circularly-averaged power spectrum (Fig. 1). An analytical function $p(f_r)$ is fitted to this spectrum, where f_r is the radial frequency (the radial coordinate in Fourier space). The pre-whitening filter is then constructed to have the circularly symmetric transfer function $w = p^{-1/2}$. Application of this filter to a micrograph then makes its noise approximately white. The removal of excess low-frequency components by the pre-whitening filter also yields a very nearly Gaussian distribution for the noise as well. It should be noted that the usual practice of high-pass filtering images before processing provides a rough sort of pre-whitening.

In the following, we assume that a data image X has been processed with a pre-whitening filter, and has been scaled such that the average pixel power is unity. That is, if the number of pixels is n_p, the total image power is

$$|X|^2 = n_p.$$

Because nearly all of the power in a typical image comes from noise, we therefore assume that the noise in such a normalized image has unity pixel variance, and so the image corresponds to the model of Eq. (1).

2.2. Matched filter

The best detector for a known signal in white, Gaussian, additive noise is the matched-filter detector (Van Trees, 1968), which is also known as the correlation detector. The input image is correlated with a noiseless reference image H to produce a correlation function C

$$C(\mathbf{r}) = \sum_{\mathbf{r}'} X(\mathbf{r}') \bullet H(\mathbf{r}' - \mathbf{r}), \tag{2}$$

and a large peak value in C indicates the presence of the desired signal (Fig. 2). The same operation can be performed as a filter in Fourier space, with the filter transfer function being the complex conjugate of the Fourier transform $\hat{H}$ of the reference:

$$C(\mathbf{r}) = F^{-1}\{\hat{X}(f)\hat{H}^*(f)\}. \tag{3}$$

If the input to the matched filter is simply zero-mean, Gaussian noise, the output correlation function C will

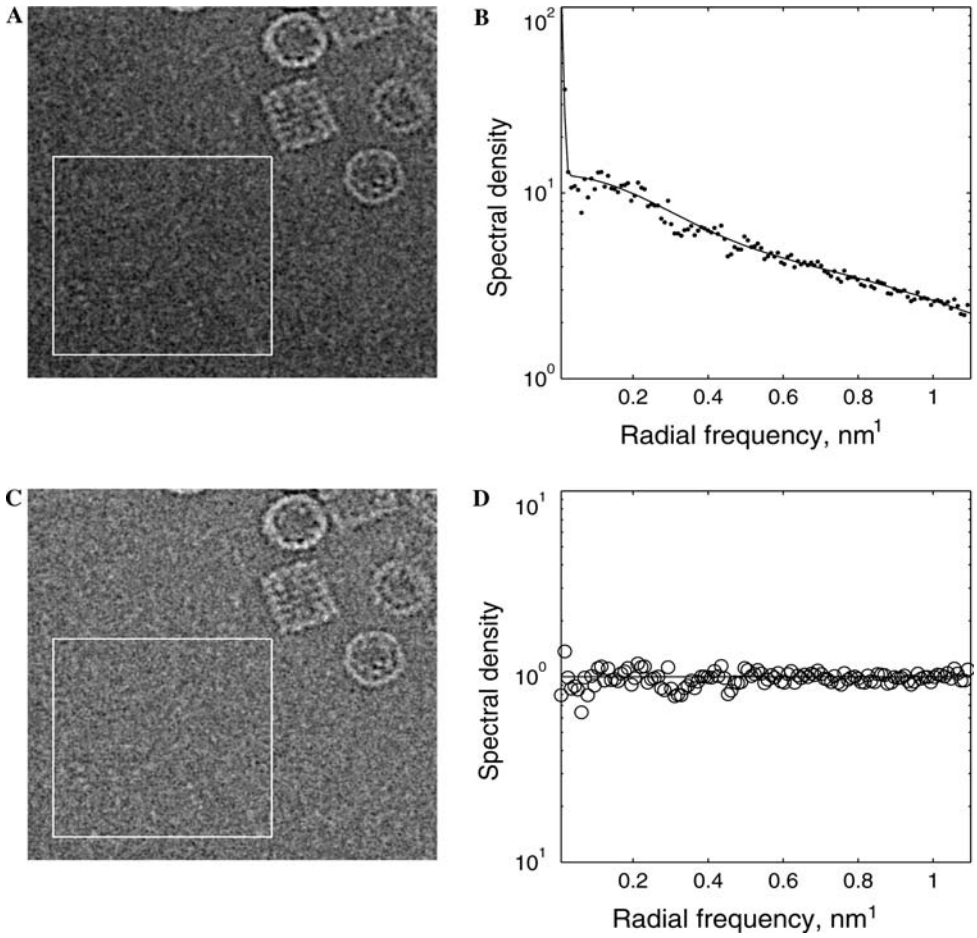

Fig. 1. Construction of the pre-whitening filter. A 210 nm square region of a micrograph is shown in (A). A particle-free region (white box) is selected and the rotationally averaged power spectrum is computed (B, points). To the mean of about 50 power spectra from the 82 images, an analytic function $p(f_r)$ is fitted (line). In this case p was the sum of three Gaussian functions, and fitting was by least-squares of log-transformed values. The pre-whitening filter frequency response is then obtained as $p^{-1/2}$ and is circularly symmetric. The pre-whitened micrograph (C) shows increased high-frequency noise and attenuated low frequencies. Its corresponding power spectrum is shown in (D).

also be zero-mean and Gaussian distributed. The matched filter is optimum in the sense that the peak value of C in response to the desired signal is maximal compared to the standard deviation of the output when only noise is present. The classical detection scheme is therefore to employ a matched filter, followed by a threshold detector to identify peaks of sufficiently large value to be regarded as reflecting true signals rather than noise.

The optimality of the matched filter detector depends on appropriate filtering of the reference. For example, the application of a pre-whitening filter to an image has the effect of standardizing the noise but at the same time modifying the underlying signal. In this case the reference image needs to be processed by the same pre-whitening filter, so that it will appropriately match the signal. Further, the matched filter principle assumes a noiseless reference. If the reference contains noise (for example, because it is derived from averaged data images) it may be appropriate to subject it to additional low-pass filtering to reduce the noise.

2.3. Density of false peaks

Even in the absence of a signal, noise alone can give rise to peaks in the correlation function. The density of false peaks exceeding a given threshold can be estimated from a result by Adler and Hasofer (1976). If white Gaussian noise is applied to the matched filter, the resulting correlation function will also consist of Gaussian noise, but now with a power spectrum proportional to $|\hat{H}(\mathbf{f})|^2$. For such a 2D Gaussian random field Adler and Hasofer found an expression for the difference in density ρ of connected regions that lie above and below the threshold value θ. We are concerned with large thresholds, in which case the regions that lie above the threshold are small and isolated. In this case ρ directly gives the number, per unit area, of false peaks arising from Gaussian noise:

$$\rho = \frac{\sqrt{2\pi}}{\sigma_0} B\theta \exp\left(-\frac{\theta^2}{2\sigma_0^2}\right). \tag{4}$$

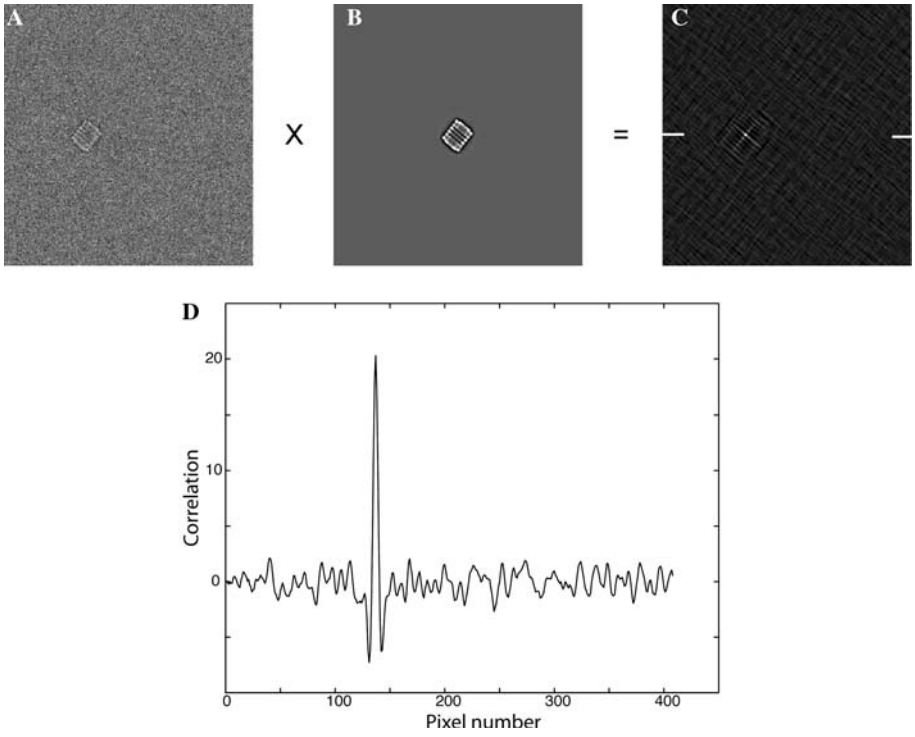

Fig. 2. Correlation (matched-filter) detector. A simulated noisy image (A) formed according to Eq. (1) with motif amplitude $m = 20$ is correlated with a reference (B) to yield the correlation function (C). A plot of pixel values (D) of the central line of the image (indicated by white markers in C) shows a sharp peak of magnitude m. Away from the position of the particle the correlation function is random and has a standard deviation of unity.

Here σ_0 is the standard deviation of the noise at the output of the matched filter and B is related to the determinant of the covariance matrix $\mathbf{A}$ of the first derivatives of the noise function by

$$B = \frac{1}{4\pi^2}|\mathbf{A}|^{1/2}.$$

In the present case, B is most easily obtained from the circularly averaged power spectrum of the noise at the filter output:

$$B = \frac{\int f_r^2 \left|\hat{H}(f_r)\right|^2 \mathrm{d}f_r}{\int \left|\hat{H}(f_r)\right|^2 \mathrm{d}f_r}, \tag{5}$$

where f_r is the radial frequency. In the case of a typical KLH particle reference (a side view, after pre-whitening) the computed value of B was $0.017\,\mathrm{nm}^{-2}$. From this value one obtains the result that, for $\theta = 4$, 5 or 6 times σ_0, the expected number of peaks exceeding threshold values to be 57, 0.8 or 0.004 per square micrometer of specimen area, respectively.

2.4. Multiple references

To search for particles in an optimum fashion, one should use not one but many different reference images corresponding to the many different orientations of the 3D particle. This overall scheme for particle detection is diagrammed in Fig. 3. For simplicity we assume here that all particle images are acquired at the same level of defocus and therefore have the same contrast-transfer function (CTF). The acquired image is processed by the pre-whitening filter and optionally by phase flipping according to the CTF (van Heel et al., 2000). The phase flipping does not improve the performance of the matched filter, but decreases the real-space dispersion (this is a potential problem when the defocus is large and the bounding box for a particle image is small) and makes particles easier to identify visually. The references must be processed in an equivalent way. If the references are computed from a CTF-corrected 3D structure, then the references need to be filtered by CTF, to correspond to the acquired images. The references must also be processed by the pre-whitening filter. Some time can be saved by noting that both the CTF and the pre-whitening filter are functions only of the radial distance from the origin in Fourier space. Thus the 3D reference structure can be filtered by 3D CTF and pre-whitening filters before projections are computed, as shown in Fig. 3A. The projections then comprise the set of references which are used by a bank of matched filters.

A large library of, say, thousands of reference images can be created by projecting an initial 3D particle structure as it is rotated about all three Euler angles. The size of this library can then be reduced by searching for the most disparate images, for example by clustering (Wong et al., 2003). I applied k-means clustering to a set

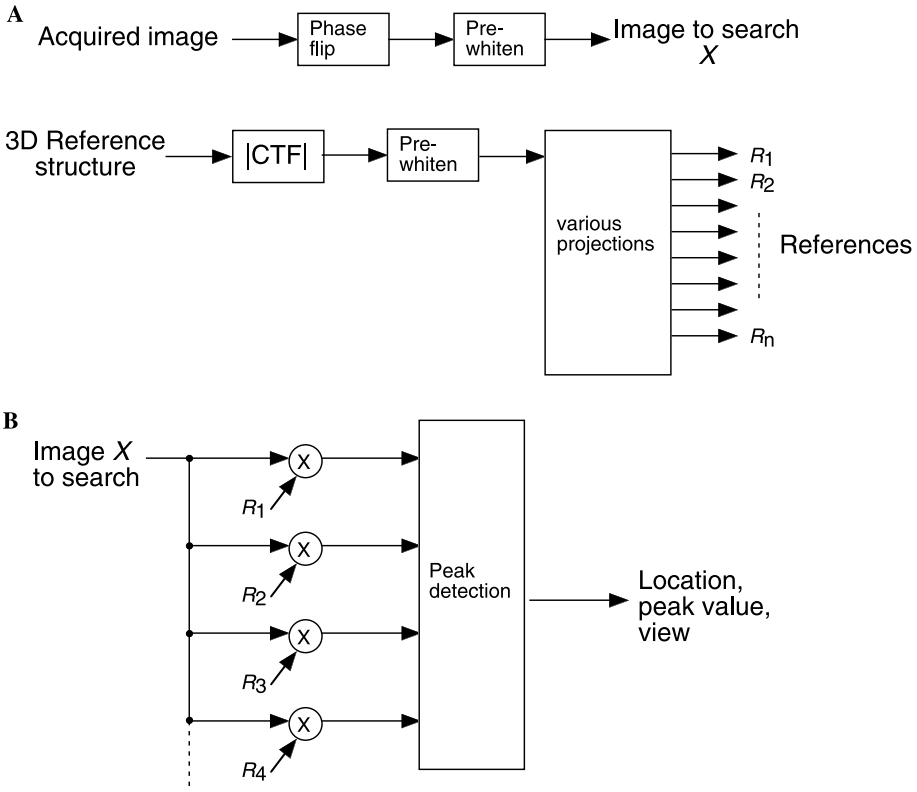

Fig. 3. Particle detection process. (A) Preprocessing. The micrograph image is processed by phase flipping (to reduce the dispersion due to the contrast-transfer function (CTF) of the microscope) and by the pre-whitening filter, and is normalized. The 3D reference structure is meanwhile filtered by the absolute value of the CTF and the pre-whitening filter. Projections are then computed from this 3D map to provide the 2D references. The result is that a motif in a reference has been processed by the same CTF and filtering operations as the same motif in the micrograph. (B) Correlation detector. The processed image is correlated with each reference, and the maximal peak values are found.

of KLH "side views" and found that rotations about the particle axis produced only small (<5%) variations in correlation values. In-plane rotations of the particle side view (Fig. 4) were much more disparate. Therefore the set of references used for the KLH dataset consisted simply of in-plane rotations of an averaged side view of the KLH particle. I used a set of 64 such images; with this number the expected peak of the correlation function is within 10% of the maximum value for any side view of the KLH particle (Fig. 4, lower panel).

It should be kept in mind that the set of reference projections will be modified by the pre-whitening filter before they are used in the matched filter detector. The effect of the pre-whitening filter may be to make some if the most disparate views more similar and vice versa. The pre-whitening filter is therefore best applied before clustering is done.

Applying the matched-filter detector with each of many different references is computationally costly. The cost can be reduced using principal component analysis to form a truncated "eigenimage expansion" of the set of references (Frank, 1996). Suppose there are n_R reference images R_i, each normalized such that $|R_i|^2 = 1$. The singular value decomposition (SVD) is employed to generate a set of orthonormal eigenimages U_j, $|U_j|^2 = 1$ and a coefficient matrix a_{ij} such that each of the R_i is given by a linear combination of the U_j,

$$R_i = \sum_{j=1}^{n_R} a_{ij} U_j \tag{6}$$

with the normalization of the R_i implying that

$$\sum_{j=1}^{n_R} a_{ij}^2 = 1 \tag{7}$$

for all i.

The eigenimages for the KLH reference set is displayed in Fig. 5. Because of linearity, the cross-correlation of a data image with a reference R_i is given by the expansion

$$C_i(\mathbf{r}) = \sum_{j=1}^{n_R} a_{ij} \Gamma_j(\mathbf{r}), \tag{8}$$

where the correlations of the data with eigenimages are

$$\Gamma_i(\mathbf{r}) = \sum_{\mathbf{r}'} \hat{X}(\mathbf{r}') \bullet U_j(\mathbf{r}' - \mathbf{r}). \tag{9}$$

In many situations the full complement of n_R terms is not required, and the expansions can be truncated to,

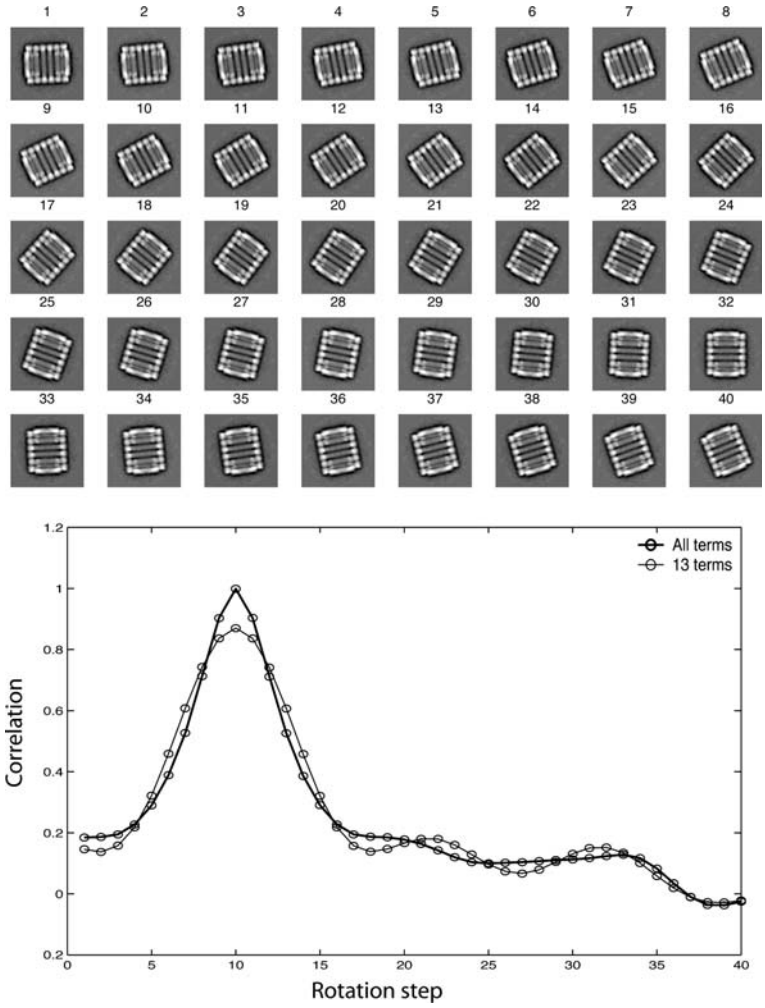

Fig. 4. Forty of the total of 64 KLH particle references. To check that the rotational degree of freedom was sampled finely enough, a correlation function (lower panel, thick curve) was computed as the inner product of reference number 10 with the others. It is seen that one step of rotation reduces the correlation by less than 10%. Also plotted as the thin curve is the correlation among references as represented in the reduced expansion with $n_t = 13$ terms.

say, n_t terms. What happens to the efficiency of detection by the matched-filter if the number of terms is reduced? We can estimate the peak-to-noise ratio at the output of the correlation detector as the number of terms is changed, as follows. Suppose the data image contains an exact copy of reference R_i, plus unity-variance pixel noise:

$$X = mR_i + N. \tag{10}$$

The expectation value of the peak output of the correlator, when n_t terms are used, will be

$$\langle C_i(0)\rangle = m\sum_{j=1}^{n_t} a_{ij}^2 (U_j \bullet U_j) = m\sum_{j=1}^{n_t} a_{ij}^2 \tag{11}$$

by the orthonormality of the U_j. Meanwhile the noise variance at the output of the correlator will be

$$\sigma^2 = \sum_{j=1}^{n} a_{ij}^2 \left\langle |U_j N|^2 \right\rangle. \tag{12}$$

The normalization of the U_j, along with the independence and unity pixel variance of the noise N, means that the expectation value

$$\left\langle |U_j N|^2 \right\rangle = 1 \tag{13}$$

for all j. Thus

$$\sigma^2 = \sum_{j=1}^{n_t} a_{ij}^2. \tag{14}$$

The peak power to noise power ratio at the output of the detector is therefore given by

$$\frac{\langle C_i(0)\rangle^2}{\sigma^2} = m^2 \sum_{j=1}^{n_t} a_{ij}^2. \tag{15}$$

This quantity is plotted as a function of the number of terms n_t in Fig. 6. It can be seen that, in the case of the KLH particles, the value $n = 13$ gives very good per-

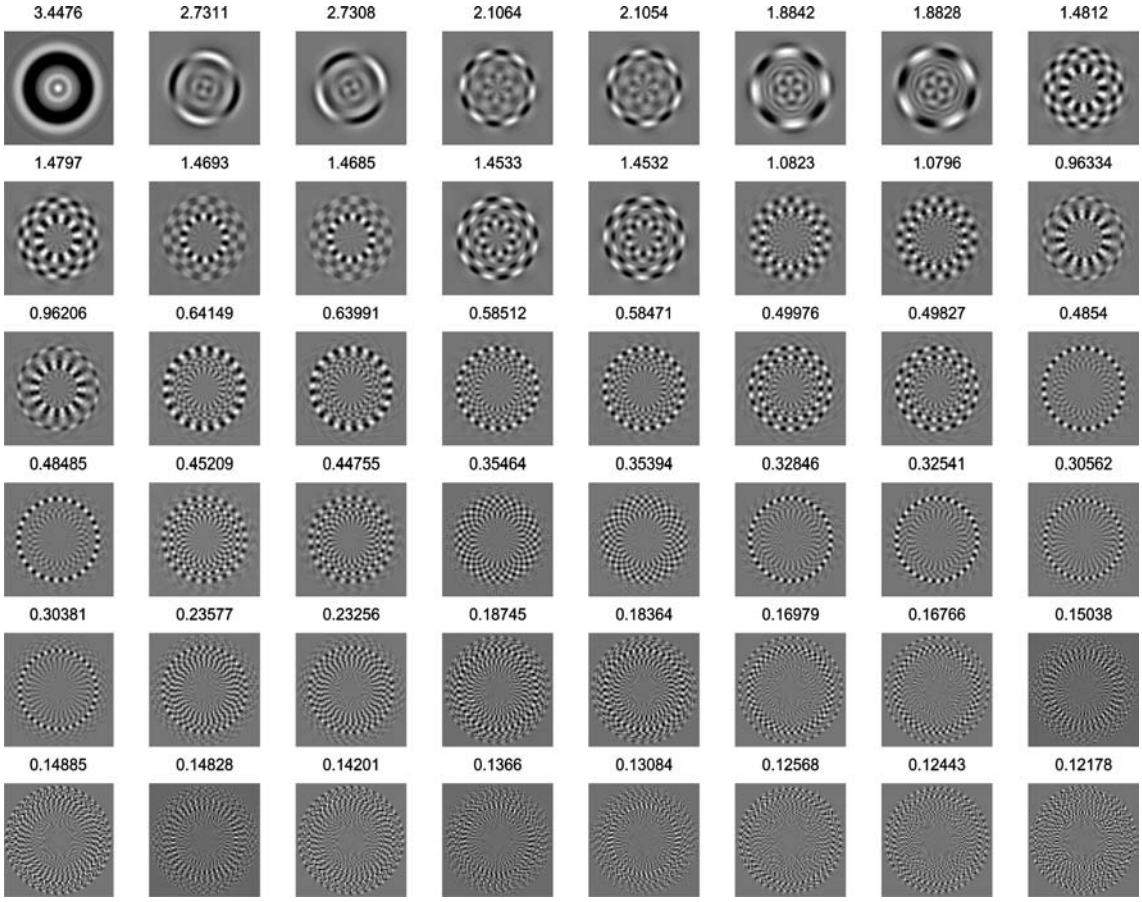

Fig. 5. Eigenimages from the principal component analysis of the references. Above each image is its singular value, which is proportional to the root mean square value of its coefficients.

formance, 90% of the value m^2 that is obtained when using all $n_R = 64$ terms.

Fig. 6 also demonstrates why, in the case of the KLH particles, a single circularly averaged image does not provide a good reference for particle detection. The first eigenimage U_1 is such a circularly averaged image, but it produces a very poor power ratio at the output of the correlator, only about 24% of the maximum.

2.5. Advantages of noise standardization

The results of the previous section illustrate the simplicity of the noise model and scaling scheme embodied in Eqs. (1) and (10). The scheme assumes that the pixel noise is independent and has unity variance. This noise is added to a particle motif A with amplitude m; but the motif A does not have unity pixel power; instead A is normalized such that $|A|^2 = 1$. The best matched filter for detecting this particle motif will be one in which the reference R_i is identical to A. In the presence of a particle the peak output of this filter has the expected value m.

In the absence of a particle, this scheme ensures that the filter's output has a standard deviation of unity. The frequency of exceeding a threshold θ due to noise will therefore be given by Eq. (4) with $\sigma_0 = 1$. Further, by inserting reasonable values into Eq. (4), we can state that for cryo-EM particles of typical size it will be difficult to discriminate particles from noise if m is smaller than about 4, but false positives due to noise will be rare if m is greater than about 8.

The most remarkable feature of this scaling scheme is that all of the properties listed above are independent of the pixel size. As long as the particle motif does not contain spatial frequencies beyond the Nyquist limit, a choice of finer or coarser sampling for the image and reference does not affect the scaling or signal-to-noise ratio at the output of matched filter, and it does not affect the frequency of false peaks.

2.6. Particle detection

We can now summarize the steps in setting up to detect particles in the presence of noise. First the noise power spectrum is evaluated, from which a pre-whitening filter is created. This filter is applied both to the data images, to give them pixel-independent noise, and to the reference structure, so that it will best match the particles in the image.

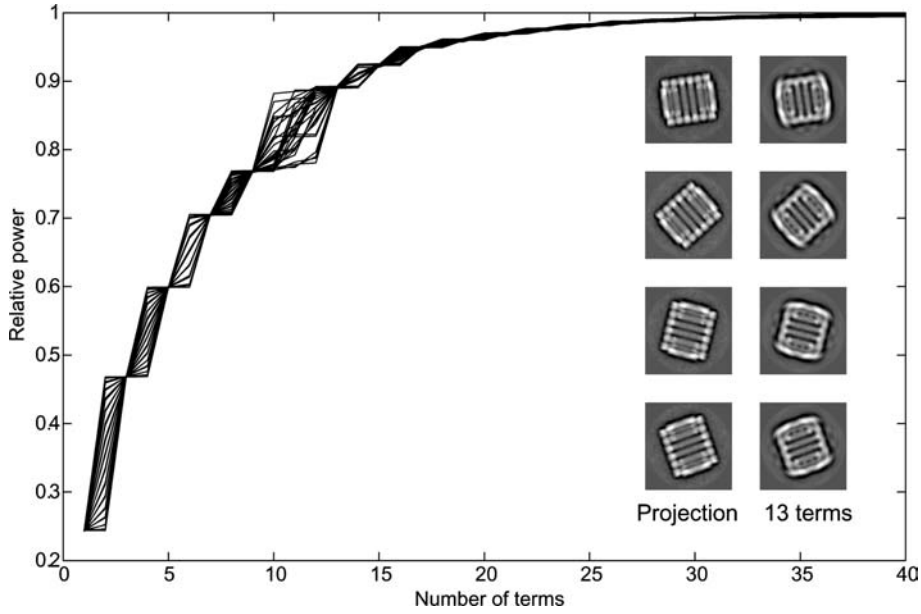

Fig. 6. Correlation detector signal-to-noise ratio (Eq. (15)) as a function of the number n_t of terms in the eigenimage expansion. Curves are drawn for each of the 64 references. I chose to use $n_t = 13$, which gives a power ratio about 90% of the maximum for all of the references. The inset compares the original references (left column) with the reconstruction from 13 terms (right column).

The second step is the creation of a representative set of reference projections. For computational efficiency an eigenimage expansion is created from this set, and a limited number of terms is chosen. At this point the correlation detector is employed.

The cross-correlations are computed between the pre-whitened data and the eigenimages. In the case of the KLH particles, 13 such correlations Γ_j are computed. For efficiency the Fourier transforms $\hat{X}_P$ of the pre-whitened data and $\hat{U}_j$ of each eigenimage is pre-computed. Their product is then inverse-transformed to yield Γ_j

$$\Gamma_j(\mathbf{r}) = F^{-1}\{\hat{X}\hat{U}_j\}.$$

The matched filter output is then approximated by

$$C_i(\mathbf{r}) \approx \sum_{j=1}^{n_t} a_{ij}\Gamma_j(\mathbf{r}). \tag{16}$$

Putative particles are identified by finding the maximum of the outputs of all the matched filters, according to the following algorithm. The first ($k = 1$) maximum s_k is found according to

$$s_k = \max_{i,\mathbf{r}}[C_i(\mathbf{r})], \tag{17}$$

while the values of i and $\mathbf{r}$ that yielded the maximum are also stored to identify the best-matching reference and the position of the putative particle,

$$(i_k, \mathbf{r}_k) = \arg\max_{i,\mathbf{r}'}[C_i(\mathbf{r})]. \tag{18}$$

Before searching for the next maximum, a circular region of the functions C_i is set to zero; this prevents double-counting of peaks. This region is centered on $\mathbf{r}_k$ and has a radius about equal to the radius of a particle. The peak search (Eqs. (17) and (18)) is repeated for $k = 2, 3, 4, \ldots$ until the identified peak value s_k falls below a threshold.

3. The particle discrimination problem

The correlation detector does not discriminate well between true particles and other objects: any image motif that provides a sufficiently large inner product will be counted as a particle. Stoschek and Hegerl (1997) have demonstrated a correlation detector that can discriminate well among different types of particles. However, in the present case where the objects to be discriminated against cannot be specified, what is needed instead is a test of similarity of the observed image to one of the references. One approach is to do a squared-error test between the particle image and a reference. A pixel-by-pixel error test is generally not practical for cryo-EM images because the particle image power is small compared to the noise power. Some kind of reduced representation of particle images is needed before an effective error test can be made.

3.1. Reduced-space representation

A pixel-by-pixel test for squared error is equivalent to evaluating the Euclidean distance in a vector space whose dimension equals the number of pixels in a particle image. Each possible image is represented by a point

in this space. One way to increase the sensitivity of the test is to eliminate most of the degrees of freedom and evaluate the distance in a reduced-dimensional space. For example, one could use the factor space of the PCA expansion. Expressing the image in terms of the orthonormal basis vectors (eigenimages) U_j the point representing a true particle image will be at a distance m from the origin, while the noise produces an uncertainty in the position consisting of one standard deviation in each of the n_d dimensions, for a total standard deviation of $\sqrt{n_d}$. In the case of the KLH images, the image amplitude $m \approx 20$; thus if n_d is fairly small, points representing true particle images will be clearly defined. I experimented with this method, but it failed to give the best results. It appears that this factor space, which is defined by the SVD to best discriminate among the various projections of the KLH particle, is ill-suited to discriminate between images of particles and non-particles.

What is needed, then, is a low-dimensional space that is capable of representing the sort of artifacts, foreign objects and overlapping particles that cause trouble for subsequent steps in single-particle processing, and which would be rejected by a human observer. After experimenting with using the space spanned by the low-order Fourier coefficients of the image, I decided to use the following strategy, which is illustrated in Fig. 7. For the kth correlation peak, the corresponding region of the pre-whitened and normalized micrograph X is extracted to form the "boxed" putative particle image W_k. The correlation detector identifies this image as being most similar to the particular reference R_{i_k}, and a least-squares estimate of the amplitude of the underlying motif is the correlation peak value s_k. The residual image, obtained by subtracting the estimated reference image and multiplying by a circular mask M, is then

$$\tilde{W}_k = (W_k - s_k R_{i_k})M. \tag{19}$$

From this residual the rotationally averaged power spectrum $p_k(f_r)$ is computed, and the statistic t is computed

$$t_k = \int_0^\infty p_k(f) w(f)\,\mathrm{d}f, \tag{20}$$

where the weighting function w is chosen to accentuate the spatial frequencies of artifacts similar in size and form to the particle. I used a weighting function of the form

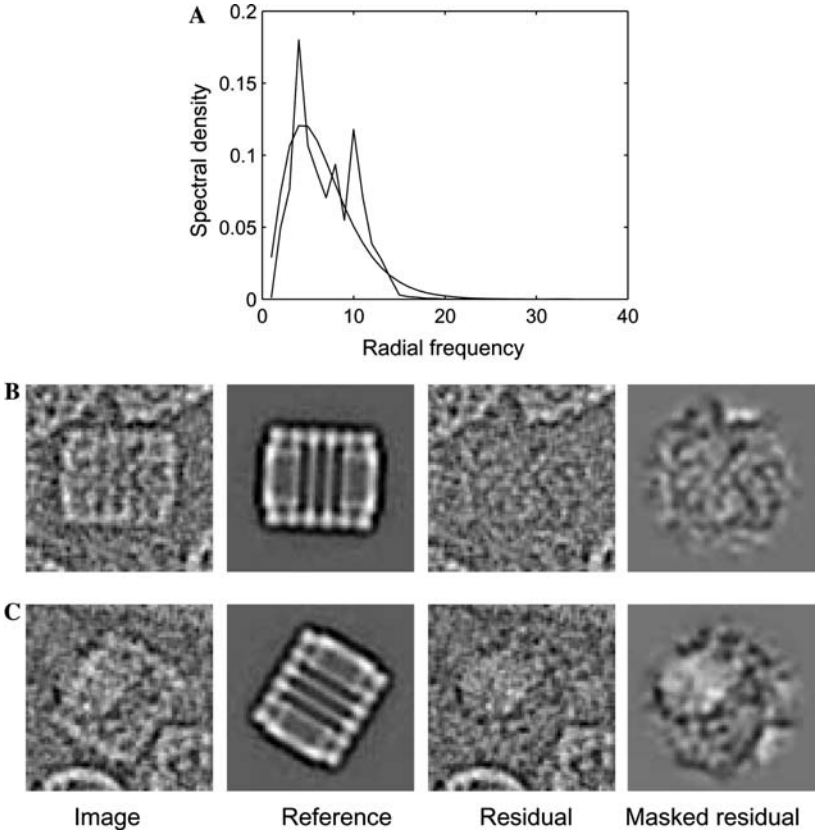

Fig. 7. Residual error calculation. (A) rotationally-averaged power spectrum of the references, and the fitted function (Eq. (21), smooth curve) which is used for weighting. The radial frequency is given in units of $(60\,\mathrm{nm})^{-1}$. (B) Construction of the residual for a "good" particle. From the values of the correlation peak the best-matching reference is scaled and subtracted to yield the residual. A circular mask is applied before computing the power spectrum of the residual. (C) The residual for a "bad" particle. For display the images were low-pass filtered (Gaussian, 0.2 times the sampling frequency f_s except for the masked residual, where the filter frequency is 0.1 f_s).

$$w(f) = \frac{b^3}{2} f^2 \mathrm{e}^{-bf}, \tag{21}$$

where the parameter b was chosen to best fit the rotationally averaged power spectrum of the references (Fig. 7A; in the case of the KLH particles, a fit yielded the value $b = 27$ nm.) This choice causes the statistic to be sensitive to the particle size but independent of the particular choice of pixel size. The expectation value of t is approximately unity, because the function w is normalized and the noise power spectrum $p(f)$ also has the expectation value one, independent of f. Indeed the expectation value would be exactly unity, were it not for a slight reduction in area caused by the circular mask. Assuming that the mask function M is equal to 1 inside a disc-shaped area and zero outside, with a total image area of n_p pixels, the assumption of white pixel noise yields the expectation value

$$\langle t \rangle = |M|^2 / n_p.$$

Because t is a sum of spectral densities, which are χ^2 distributed random variables, its distribution can be computed numerically, or by a simple Monte-Carlo simulation.

3.2. Particle discrimination in practice

The two statistics s and t can be used to identify particles. At high signal-to-noise ratio the expectation value of s equals the motif amplitude m in the image model (Eq. (1)); however at low signal-to-noise some values of s may represent random noise peaks. A minimum acceptable value for s can be set on the basis of the allowed frequency of false positives arising from noise, as estimated from Eq. (4).

For identical particles in uniform noise, the values of s are expected to have a Gaussian spread. Because of the standardization of the noise and the normalization of the references, the standard deviation of this Gaussian will be unity. In practice (Fig. 8A) s takes on a broader distribution as the thickness of ice and other imaging factors cause some particles to be imaged with higher amplitude than others. In the case of the KLH dataset, s values for "good" particles showed a broad distribution ranging from about 14–22. False particles due to noise peaks are expected to have s values of about 6 or less.

The statistic t is more robust in the sense that its distribution is almost entirely determined by the properties of the background noise. The observed distribution (Fig. 8B) is quite close to theory, and t can serve as a strong criterion for distinguishing "good" from "bad" particles.

The use of these two statistics is illustrated in the scatterplot of Fig. 9. Each putative particle is represented by a point in the s–t plane. The "good" particles

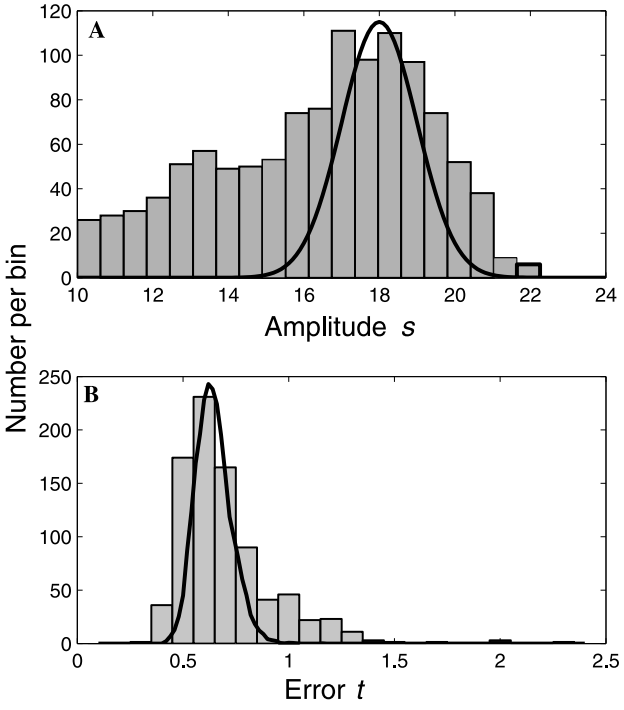

Fig. 8. Distribution of the statistics s and t in the KLH dataset. (A) The distribution of s is much broader than the Gaussian, unity SD distribution predicted by the noise model (solid curve). This arises from variations in motif amplitude among individual micrographs in the dataset. (B) The distribution of the squared error t lies close to the theoretical distribution, computed from the spectral weighting function (Eq. (17)) and assuming unity spectral density of the noise. In (A), the histogram of s was computed for all correlation peaks for which $t < 1.3$. In (B), the histogram of t is for all $s > 14.5$.

were identified by matching their image coordinates with a set of manual picks from the same dataset. The "bad" particles are all putative particles that could not be matched with a manually picked one. Most good particles are seen to be clustered in a region that is bounded by a minimum value of s and a maximum value of t. More importantly, this region encloses only a small number of bad particles. The manual picks identified 951 "good" particles in the dataset. Of these, 838 lay within the chosen region of the s–t plane, along with 38 "bad" particles. These numbers correspond to a false-negative rate of 12% and a false-positive rate of only 4%.

The algorithm presented here was implemented in the Matlab environment (MathWorks, Natick, MA) using standard built-in functions including SVD for the singular value decomposition and FFTN for Fourier transforms and filtering. The time to perform the correlations and peak searches for each micrograph (binned to 512×512 pixels) required about 20 s of computation time on a 2 GHz personal computer. This being an experimental study, a stand-alone program was not built. However, the Matlab code is available on request from the author.

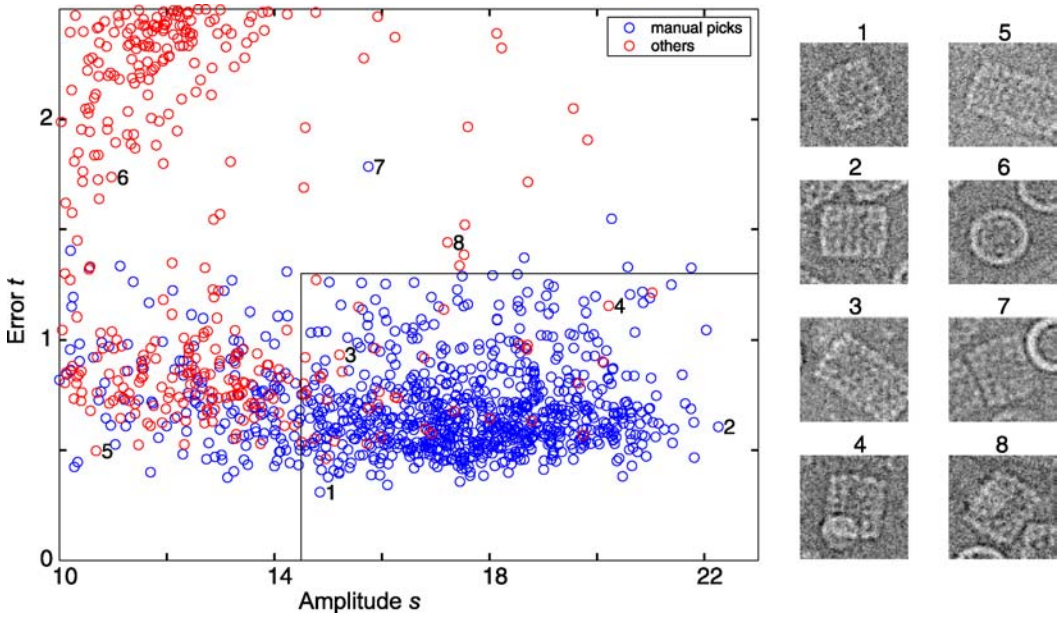

Fig. 9. Clusters of particles identified by the two statistics s and t. Coordinates of "good" manually picked particles were matched with those of correlation peaks and are plotted as blue points; all other putative particles are plotted in red. It is seen that the region defined by $s > 14.5$ and $t < 1.4$ encloses nearly all "good" particles. On the right individual particle images are shown that correspond to the numbered points. The first column are images that would be marked "good" by the algorithm; the second column are ones that are marked "bad."

4. Discussion

Described here is an automatic particle selection algorithm that shows good performance on a simple dataset having an excellent signal-to-noise ratio. Its performance on more challenging datasets has not been tested, but there is hope that some of the underlying principles—a standardized noise model, a fast algorithm for multiple correlations, and a discrimination statistic t with a predictable distribution—may be combined with other approaches to yield a truly robust automatic particle selector.

4.1. Free parameters

A model-based automatic particle-picking algorithm would ideally require very little input besides the preliminary 3D model of the particle and information about the contrast-transfer function. A goal in the implementation described here was to minimize the number of parameters that the user would need to set. Not counting the parameters that affect the processing speed but have small effects on the quality of particle selection (for example, the pixel size and the number of references) the only parameters that were specified in this example were the approximate size of the particle (for construction of a mask) and the bounds for the statistics s and t. In the end, the only truly arbitrary parameter was the bound for s. A guess of the particle size was used to set up the circular mask; however much better masks could have been derived automatically from a 3D model. The theoretical distribution for t can be used to select a suitable upper bound for its values. An appropriate bound for s however cannot be determined as easily. It should be large enough to guarantee an acceptably low-frequency of spurious noise peaks, but otherwise its value must be determined empirically, for example through histograms as in Fig. 8A. In the case of the KLH particles it is seen that an appropriate threshold for s was about 14, far above the value of 6 which would be constrained by an acceptable noise-peak rate. The larger value was needed in order to discriminate against incomplete or distorted particles as can be seen in Fig. 9.

The dataset considered here was particularly simple because an in-plane rotation gave a complete set of references. Were we instead trying to accept the very different views of an asymmetric 3D particle, setting bounds for the statistic s would be more complicated. This is because different projections of the same particle can have different amounts of image power, and therefore different values of the amplitude m in the image model (Eq. (1)). Since the expectation value of s is equal to m, threshold values of s will vary depending on which view is being detected. However, one can imagine a simple scheme in which the threshold values for the s_k can be set by a simple procedure. Each threshold must above a level determined by the allowable noise-peak density (Eq. (4)), but otherwise would be proportional to the motif amplitude of each projection.

4.2. Assumption of uniform noise

A critical assumption in the algorithm described here was that the micrographs in the dataset have uniform and identical noise statistics. For the construction of the pre-whitening filter, an average power spectrum from the entire set of micrographs was employed. In reality, individual micrographs, and even different areas within a micrograph, have differing proportions of low-frequency noise to the background shot noise. Indeed, the non-ideal distributions of s and t seen in Fig. 8 are readily explained by heterogeneity in the micrographs. Variations in ice thickness yield variations in the motif amplitude while the shot noise remains essentially unchanged. This gives a broad variation in the apparent motif amplitude s.

Variations in the amplitude of low-frequency noise tend to skew the distribution of t toward higher values, as was observed. Thus an improvement in the algorithm would result from the ability to characterize the local noise spectrum within an image. A compensation for the local variance is effectively provided in the local correlation procedure of Roseman (2003), and in the likelihood ratio test of Wong et al. (2003). An extension of these corrections by treating low and high frequency noise components separately could result in a more robust algorithm.

4.3. The advantage of multiple templates

The increasing speed of computers makes it quite practical now to employ multiple-reference searches for particles. The value of using multiple references is made clear by the result of Eq. (15) and Fig. 6, which show how the detection sensitivity varies with the size of the eigenimage expansion. In the case of the KLH particles, a reference consisting only of the rotational average gave less than 25% of the maximum signal-to-noise ratio at the output of the correlation detector. The poor performance of the single rotational average can be understood from the non-globular shape of the KLH particle and also from the reduction in low-frequency components due to the contrast-transfer function and from the pre-whitening filter. It should be kept in mind that the attenuation of low frequencies by the pre-whitening filter is necessary to prevent low-frequency noise and artifacts from producing false positives. The result however is that higher-frequency components become more important in the particle recognition process, and higher-order terms and multiple references are required to efficiently identify particles.

Acknowledgments

I am grateful to Shirley Wang (Yale College) for assistance in programming. I also thank Professors Peter Schultheiss (Yale), Eric Hansen (Thayer School of Engineering, Dartmouth College), and Marshall Bern (Palo Alto Research Center) for advice and discussions. The data used here were provided by the National Resource for Automated Molecular Microscopy (supported by National Center for Research Resources Grant No. RR17573). The author's work was supported by NIH Grant No. NS21501.

References

Adler, R.J., Hasofer, A.M., 1976. Level crossings for random fields. Ann. Probability 4, 1–12.

Frank, J., 1996. Three Dimensional Electron Microscopy of Macromolecular Assemblies. Academic Press, New York.

Frank, J., Wagenknecht, T., 1984. Automatic selection of molecular images from electron micrographs. Ultramicroscopy 12, 169–176.

Lata, K.R., Penczek, P., Frank, J., 1995. Automatic particle picking from electron micrographs. Ultramicroscopy 58, 381–391.

Ludtke, S.J., Baldwin, P.R., Chiu, W., 1999. EMAN: semi-automated software for high resolution single particle reconstruction. J. Struct. Biol. 128, 82–97.

Nicholson, W.V., Glaeser, R.M., 2001. Automatic particle detection in electron microscopy. J. Struct. Biol. 133, 90–101.

Ogura, T., Sato, C., 2003. Automatic particle pickup method using a neural network has high accuracy by applying an initial weight derived from eigenimages: a new reference free method for single-particle analysis. J. Struct. Biol. 145, 63–75.

Roseman, A.M., 2003. Particle finding in electron micrographs using a fast local correlation algorithm. Ultramicroscopy 94, 225–236.

Stoschek, A., Hegerl, R., 1997. Automated detection of macromolecules from electron micrographs using advanced filter techniques. J. Microsc. 185, 75–84.

Wong, H.C., Chen, J.D., Mouche, F., Roullier, I., Bern, M., 2003. Model-based particle picking for cryo-electron microscopy. J. Struct. Biol. 145, 157–167.

van Heel, M., Gowen, B., Matadeen, R., Orlova, E.V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M., Patwardhan, A., 2000. Single-particle electron cryo-microscopy: towards atomic resolution. Q. Rev. Biophys. 33, 307–369.

Van Trees, H.L., 1968. Detection, Estimation, and Modulation Theory. Wiley, New York. 697 pp.

Zhu, Y., Carragher, B., Mouche, F., Potter, C.S., 2003. Automatic particle detection through efficient Hough transforms. IEEE Trans. Med. Imaging 22, 1053–1062.