BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 19 2013, pages 2460–2468
doi:10.1093/bioinformatics/btt429

Bioimage informatics

Advance Access publication August 19, 2013

A pattern matching approach to the automatic selection of particles from low-contrast electron micrographs

V. Abrishami[1,2,*], A. Zaldívar-Peraza[1,2], J. M. de la Rosa-Trevín[1,2], J. Vargas[1], J. Otón[1], R. Marabini[2], Y. Shkolnisky[3], J. M. Carazo[1] and C. O. S. Sorzano[1,4,*]

[1]Biocomputing Unit, National Center of Biotechnology (CSIC), [2]Department of Computer Science, University Autonoma de Madrid, Campus Universidad Autonoma s/n, 28049 Cantoblanco, Madrid, Spain, [3]Department Applied Mathematics, Tel Aviv University, Ramat Aviv, Tel Aviv 69978 Israel and [4]Bioengineering Lab, Escuela Politecnica Superior, University San Pablo CEU, 28668 Boadilla del Monte, Madrid, Spain

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Structural information of macromolecular complexes provides key insights into the way they carry out their biological functions. Achieving high-resolution structural details with electron microscopy requires the identification of a large number (up to hundreds of thousands) of single particles from electron micrographs, which is a laborious task if it has to be manually done and constitutes a hurdle towards high-throughput. Automatic particle selection in micrographs is far from being settled and new and more robust algorithms are required to reduce the number of false positives and false negatives.

Results: In this article, we introduce an automatic particle picker that learns from the user the kind of particles he is interested in. Particle candidates are quickly and robustly classified as particles or non-particles. A number of new discriminative shape-related features as well as some statistical description of the image grey intensities are used to train two support vector machine classifiers. Experimental results demonstrate that the proposed method: (i) has a considerably low computational complexity and (ii) provides results better or comparable with previously reported methods at a fraction of their computing time.

Availability: The algorithm is fully implemented in the open-source Xmipp package and downloadable from http://xmipp.cnb.csic.es.

Contact: vabrishami@cnb.csic.es or coss@cnb.csic.es

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 26, 2013; revised on June 28, 2013; accepted on July 19, 2013

1 INTRODUCTION

Electron Microscopy (EM) is a key tool to study the structure and function of biological macromolecules at a medium–high resolution. Single particle analysis is an EM modality in which multiple copies of the same macromolecule are simultaneously imaged into a single micrograph. Particles from several hundreds or even thousands of micrographs are commonly employed in a structural study. The standard data processing workflow of single particle reconstruction includes: particle selection, particle alignment, particle classification, three-dimensional (3D) reconstruction and model refinement (Chen and Grigorieff, 2007; Sorzano *et al.*, 2012). Different views of a specimen are required for the 3D reconstruction of a complex, but these views suffer from low signal-to-noise ratio (SNR) (due to low-dose imaging) (Glaeser, 1971), low contrast (due to close to focus conditions), and image deformations (due to the microscope aberrations). It is generally accepted that high resolution can only be achieved with thousands of projection images, so that the 3D reconstruction algorithm can compensate for these challenging imaging conditions. In particular, there is generally a direct relationship between the number of selected particles and the maximum achievable resolution (Henderson, 1995). Manually identifying that number of particles is not just time consuming and laborious, but also an error-prone process. A robust automatic particle picker (APP) algorithm is, therefore, indispensable to enhance the technique's throughput.

As the selection of several thousands of particles from low-dose micrographs is the first and one of the crucial steps towards a high-resolution reconstruction, a large amount of effort has been made by researchers to develop accurate methods for APP. These methods have been classified into groups by different authors (Mallick *et al.*, 2004; Nicholson and Glaeser, 2001; Zhu *et al.*, 2004). Among them, the classification by Nicholson and Glaeser (2001), suggests a general categorization into template matching-based and feature-based approaches. Template matching-based methods (Chen and Grigorieff, 2007; Huang and Penczek, 2004; Ludtke *et al.*, 1999; Plaisier *et al.*, 2004; Roseman, 2003; Sigworth, 2004; Wong *et al.*, 2004) calculate the cross-correlation (or any other measure of similarity) between a set of templates and a micrograph image to seek for particle candidates. Templates are obtained either from different projections of an initial 3D volume (Huang and Penczek, 2004; Wong *et al.*, 2004) or from a number of manually picked particles (Hall and Patwardhan, 2004; Roseman, 2003). Instead of using all templates (either from different projections of an initial 3D volume or from a number of manually picked particles), which severely increases the processing time, other alternatives can be employed, such as eigenimages of templates (Sigworth, 2004) or some form of an average of each template cluster (Wong *et al.*, 2004). In feature-based approaches, particles are sought through the calculation of some prominent geometric and/or statistical features of the particle images (Arbeláez *et al.*, 2011; Hall and Patwardhan, 2004; Langlois *et al.*, 2011; Mallick

*To whom correspondence should be addressed.

et al., 2004; Ogura and Sato, 2004; Sorzano *et al.*, 2009; Volkmann, 2004; Yu and Bajaj, 2004; Zhao *et al.*, 2013; Zhu *et al.*, 2003). Feature-based methods can be reference-free or learning-based. In the former, features corresponding to particles are known to fall within a certain region of the feature space and, therefore, no training is necessary and the algorithm can start picking particles straightaway; in the latter, however, a set of particles and non-particles are required to train a classifier which is then able to distinguish between particles and non-particles based on the training features. Although reference-free methods require less effort from the user, they are of limited applicability, because the space region corresponding to particle features has to be known *a priori*.

We previously introduced a feature-based APP method (Sorzano *et al.*, 2009) that learns features from the user selected particles via a continuous learning phase (the algorithm is available in the open-source package Xmipp 2.4). In a manual picking step, a small dataset of particle and non-particle images is formed to train an ensemble naive Bayesian classifier. Once the classifier is trained, it suggests new particles in a new micrograph. The user supervises this result by discarding the wrongly picked particles and identifying the disregarded ones. This feedback information is then submitted to the classifier which is updated to accommodate this new information. This semi-automatic picking is continued on several micrographs until the user is satisfied by the results. At this point, the trained classifier carries out the selection of the particles in the remaining micrographs in a fully automatic way.

In this article, we introduce an APP method that follows the general learning structure of Sorzano *et al.* (2009), but major improvements are made to increase speed and accuracy. Our new feature vector is completely different from the previous method and consists of a number of geometrical and statistical features; it is robust to noise, very fast to compute and most of its features are rotationally invariant. Instead of Naive Bayesian (NB), we now use a support vector machine (SVM) as the base classifier due to its interesting properties, like high generalization capabilities and small training/classification time. In order to reduce the number of false positives, two SVM classifiers are used: one for discriminating between particle candidates and non-particle objects and the other for checking if a particle candidate recognized by the first classifier is a real particle or not. In contrast to Sorzano *et al.* (2009), which explores a big search space for particles, the proposed method limits the search space to the peaks obtained from the cross-correlation of the micrograph with some pre-computed templates. Templates are generated during the manual picking step by clustering [the clustering algorithm is described in Sorzano *et al.* (2010)] the hand-picked particles and selecting the average of each cluster. As the correlation with all orientations of the template noticeably increases the computation time, a rotationally averaged template can be used instead. The algorithm has been successfully tested on three experimental datasets and compared with our previous algorithm (which had, in turn, been compared with other approaches), resulting in a more than an order of magnitude decrease in computing time while achieving even better performance.

2 METHODS

There are three crucial steps in the proposed algorithm: identifying initial possible locations of particles, locally characterizing the image by means of a low-dimensional feature vector and classifying each location as particle or non-particle. In this section, we go through the details of each step.

2.1 Identification of possible particle centres

In principle, each pixel of a micrograph has the potential to be the centre of a particle. In practice, checking each pixel is rather time consuming, and so the number of possible candidates to evaluate must be reduced. Therefore, we cross-correlate the input micrograph with suitable templates to make an initial guess about the candidate particle positions. In this way, not only the search space is noticeably reduced, but also the rate of false positives. During the training phase, we cluster manually picked up particles into a few number of classes using an algorithm similar to the one described in Sorzano *et al.* (2010). The number of clusters can be increased or decreased dynamically to facilitate user's ongoing control on the quality of the templates. These templates are correlated with the input micrograph at all possible orientations (at each location of the correlation map, we should keep the maximum observed correlation for all templates and all orientations). If the particle is relatively globular, the calculation of the correlation map can be further accelerated by substituting the template by its rotational average, avoiding in this way the need to correlate with all possible orientations.

Local maxima of the correlation map are possible candidate locations for being at the centre of a particle. We further reduce this candidate set by ranking local maxima according to their correlation values and keeping only a portion of the local maxima with a value higher than a small threshold. This portion may not be the same for all datasets, and basically depends on the density of particles within the micrographs.

We should note that the cross-correlation with a template in our approach is just a first step to have a set of possible candidates, but that in no way do we rely on this cross-correlation for further processing of these candidates into particles and non-particles. In fact, the chosen threshold for local maxima is small enough such that even particles that were related to excluded templates could still be detected at this initial step (Fig. 1). In most practical cases we have found that the use of only one template is generally enough. For instance, in Figure 1, only a single class representative has been azimuthally averaged and cross-correlated with the micrograph, to identify the local maxima. In this figure, picking

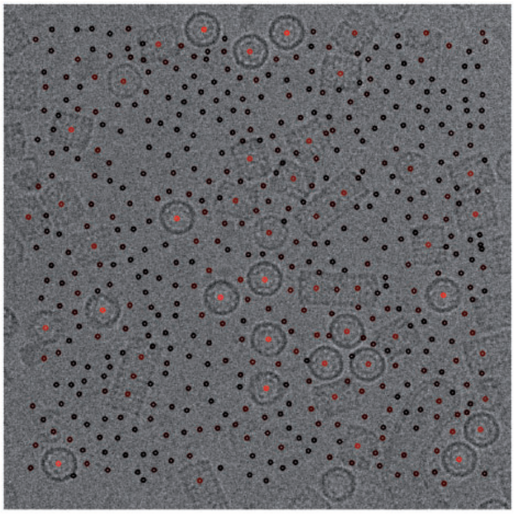

Fig. 1. Micrograph from the KLH training dataset with cross-correlation peaks superimposed. The peaks have been obtained by cross-correlation of a single class representative and the micrograph. Colours show the energy of each peak by black as lowest and red as highest

10% of local maxima (with the highest cross-correlation values) results in local maxima with minimum value of 0.27 and maximum value of 0.98. As can be seen, at the location of each particle a local maximum with a specific energy can be observed. Still, this behaviour depends on the practical case at hand, and consequently, the number of templates is dynamic, so if the result is not satisfactory the user can request for more templates.

2.2 Feature extraction

We compute a feature vector at each location previously identified as a candidate for particle centre. This vector is used by the classifier in the next step to distinguish between particles and non-particles. Two properties have been sought for the feature vector: being robust to noise (due to the low SNR in the micrographs) and being rotationally invariant (to save computational time and avoid having to look for the particles in all possible orientations). We use a feature vector that is robust to noise and most of its features are rotationally invariant. it is made of three subsets of features: the first two feature subsets are sensitive to the particle shape, whereas the third one encodes the particle grey intensities. The first and third feature subsets are rotational invariant but not the second one.

2.2.1 Particle shape description at different frequencies Micrographs are submitted to a filter bank with N_h raised cosine band pass filters to decompose them into several sub-band images in order to being able later on to extract features associated to particular frequencies (Fig. 2). The filtered micrograph by the kth ($k = 1, \ldots, N_h$) filter, $M_k(\mathbf{r})$ (i.e. kth sub-band image) is computed by

$$M_k(\mathbf{r}) = FT^{-1}\{M(\mathbf{R})H_k(\mathbf{R})\}, \tag{1}$$

where FT^{-1} is the inverse Fourier transform and $M(\mathbf{R})$ and $H_k(\mathbf{R})$ show the Fourier transform of the micrograph and kth Fourier filter in the filter bank, respectively. $\mathbf{R}$ is the two-dimensional (2D) spatial frequency, and $\mathbf{r}$ is the spatial coordinate within the image. Each raised cosine band pass filter has a particular width Δ_R and a decay of δ_R (see the last image in column B of Fig. 2). The low-pass filter in the filter bank is defined by a transfer function given by

$$H_1(\mathbf{R}) = \begin{cases} 1, & R \le \Delta_R \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi(R-\Delta_R)}{\delta_R}\right)\right), & \Delta_R < R \le \Delta_R + \delta_R \\ 0, & R > \Delta_R + \delta_R \end{cases} \tag{2}$$

where R is the modulus of $\mathbf{R}$. The kth band pass filter ($k = 2, 3, \ldots, N_h$) is defined by the transfer function $H_k(\mathbf{R})$ defined as

$$\begin{cases} 0, & R \le (k-1)\Delta_R - \delta_R \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi(R-(k-1)\Delta_R)}{\delta_R}\right)\right), & (k-1)\Delta_R - \delta_R < R \le (k-1)\Delta_R \\ 1, & (k-1)\Delta_R < R \le k\Delta_R \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi(R-k\Delta_R)}{\delta_R}\right)\right), & k\Delta_R < R \le k\Delta_R + \delta_R \\ 0, & R > k\Delta_R + \delta_R \end{cases} \tag{3}$$

Let us concentrate now on a given particle within a micrograph. We will refer to this boxed image as $I(\mathbf{r})$. Let us call $I_k(\mathbf{r})$ the corresponding boxed image extracted from $M_k(\mathbf{r})$. By considering r as the distance from the image centre in the direction of θ, we express this boxed, filtered image in polar coordinates, $I_k(r, \theta)$ and compute the cross-correlation function of pairs of $I_k(r, \theta)$ images [this step has been partially inspired by Schatz and van Heel (1990)]. Note that the cross-correlation is rotationally invariant in 2D polar coordinates (see proof in Supplementary material). Let us define $\psi_{kk'}(\Delta r, \Delta\theta)$ as the cross-correlation function between polar, sub-band images $I_k(r, \theta)$ and $I_{k'}(r, \theta)$, as below

$$\psi_{kk'}(\Delta r, \Delta\theta) = \sum_r \sum_\theta I_k(r, \theta) I_{k'}(r + \Delta r, \theta + \Delta\theta), \tag{4}$$

where $I_{k'}(r, \theta)$ is shifted over $I_k(r, \theta)$ by Δr ([0, r_p], where r_p is the radius of the particle), and by $\Delta\theta$ ($[-180^{\circ}, 180^{\circ}]$) along r and θ, respectively (see column D in Fig. 2). In particular, for each k, we calculate $\psi_{kk'}$ for $k' = k$ (autocorrelation; $k = 1, \ldots, N_h$), $k' = k + 1$ ($k = 1, \ldots, N_h - 1$) and $k' = k + 2$ ($k = 1, \ldots, N_h - 2$). The autocorrelation of a given band is related to the particle shape, and the cross-correlation between sub-bands reveals the linear relationships between the shapes at two different frequency bands. We can see in Figure 3 an example of how these cross-correlation functions can, indeed, distinguish between particles and non-particles. From this figure, it is clear that cross-correlation functions for particles (first row) are quite different from those for non-particles (second row). It is worth mentioning that, for the sake of speed up, we just consider the 2D projections and not the original 3D object. According to this, cross-correlation functions are not rotationally invariant in 3D, and therefore we need to have enough projections from different orientations to fully cover the projection space.

The cross-correlation functions of the training set of particles for a particular combination kk' are highly redundant and can be easily compressed using principal component analysis (PCA) (Pearson, 1901). There are as many $\psi_{kk'}$ images as candidate particles, which is a number in the many thousands. This is the set from which the PCA basis is extracted (there are, therefore, as many PCA's as all combinations of k and k'). For each $\psi_{kk'}$ image, we keep its projection onto the first N_b PCA vectors as features to be used during the classification step. Note that the PCA basis is calculated for the kk' cross-correlations of the training dataset of particles. This means that the kk' cross-correlations of non-particles will be poorly represented by this PCA basis. In this way, the dimensionality reduction itself is presumed to have a positive impact on the classification accuracy. An example of these bases can be seen in Figure 4, where four eigenvectors are shown for each cross-correlation function.

2.2.2 Particle shape description in a particle-adapted rotational invariant subspace Ponce and Singer (2011) proposed to calculate an image basis that is adapted to the kind of images being studied and their in-plane rotations. They do this by calculating the PCA of a set of images and all their possible rotations. We apply this principle to extract some features from boxed particles or particle candidates [this idea is partially the same as in Dube *et al.* (1993)]. Given the training particles provided by the user, we first align them into a few templates using a process similar to that described in Sorzano *et al.* (2010). Then, we compute the PCA basis associated to these templates (Fig. 5) to form a rotational invariant subspace (the basis of this subspace is able to reproduce any 2D rotation of particle images). We keep the first N_{rb} projection coefficients onto this basis as part of the feature vector. Note that the subspace that is spanned by these vectors is rotationally invariant, but not the basis itself. Therefore, particles with different 2D orientations have different coefficients but still can be efficiently approximated using this basis. Again, this basis has been especially designed to represent good particle images; consequently, non-particles will be poorly represented by the basis helping the classifier to perform its task.

2.2.3 Particle intensity Desired particles are assumed to follow a specified pattern of intensity distribution once the boxed image is normalized to have zero mean and unit power (Sorzano *et al.*, 2004). For instance, particles in the carbon region are normally discarded. To capture the intensity features desired by the user we calculate the mean, the SD as well as N_i equidistributed deciles of the intensity histogram.

2.2.4 Feature vector We collect all these features into a feature vector that characterizes the boxed image to be classified. The vector size is $3N_b(N_h - 1) + N_{rb} + N_i + 2$. By default, we suggest to use $N_b = 4$, $N_h = 6$ ($\Delta_R = 0.025$, $\delta_R = 0.02$ in digital frequencies normalized to 0.5), $N_{rb} = 20$ and $N_i = 9$. This produces a feature vector of dimension 91. In practice, we have observed that these choices provide generally good results on all the tested datasets. The values of the parameters depend on

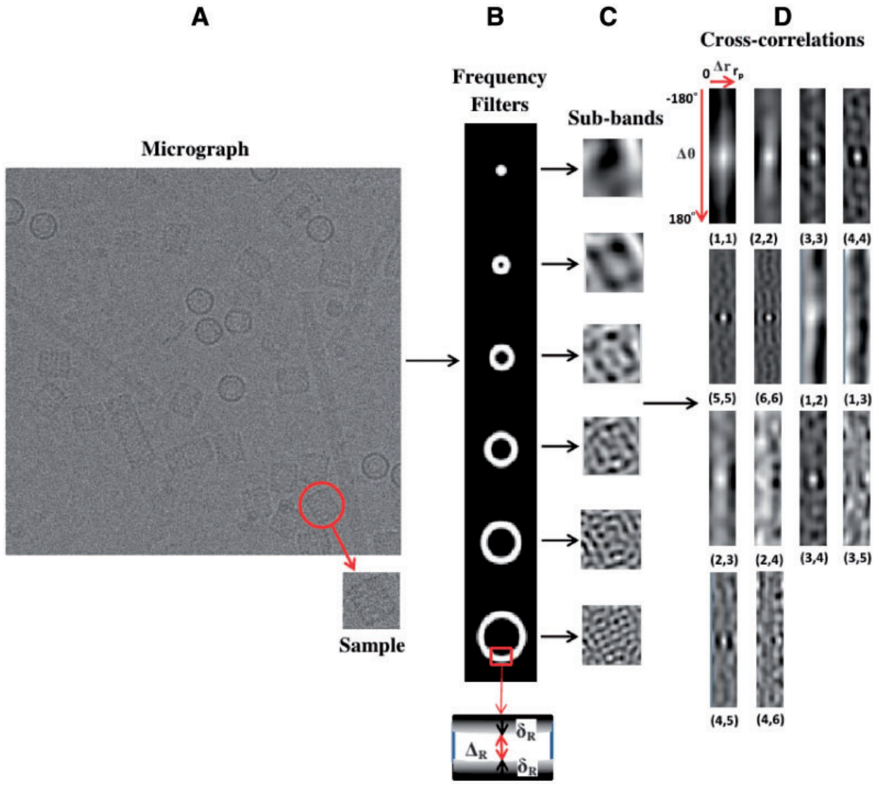

Fig. 2. Particle shape description at different frequencies. First, N_h bandpass filters (column B), with a width of Δ_R and a decay of δ_R (see the last image in this column), are applied to the input micrograph (column A) and boxed images are extracted at the location of particles or particle candidates (column C). Sub-band images for the particle (column C) are converted to polar form, and cross-correlations between different sub-bands are calculated (column D). $\Delta\theta$ $[(-180^{\circ}, 180^{\circ})]$ and Δr ($[0, r_p]$, where r_p is the radius of particle) are shifting parameters to slide one polar image on one another. The indexes below each cross-correlation (column D) refer to the polar images that form it. Parameters for this figure are $N_h = 6,\ \Delta_R = 0.025,\ \delta_R = 0.02,\ r_p = 25$ (see the main text)

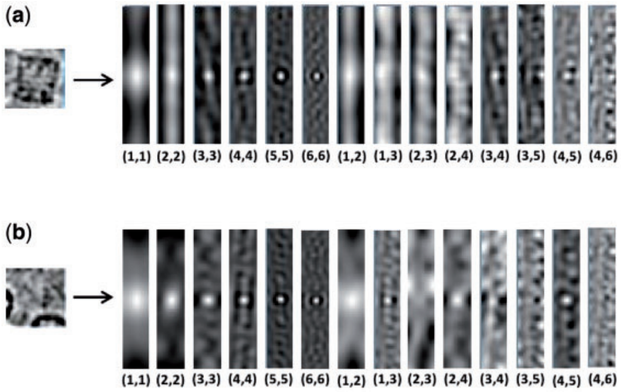

Fig. 3. Particle shape description at different frequency bands: cross-correlation among sub-bands for a particle (**a**) and non-particle (**b**). The indexes below each cross-correlation refer to the k and k' sub-bands. Note the important differences between particles and non-particles for the functions (1, 1), (2, 2), (3, 3), (1, 2), (1, 3), (2, 3) and (2, 4)

both properties of the micrographs and particles. For instance, for a low defocus set of micrographs, we need a higher value of N_h to extract the information of higher frequencies, whereas for smaller particles we should increase N_b and N_{rb} to be able to regenerate them from the basis.

2.3 Classification

Distinguishing between particles and non-particles is, in general, a complex task due to the low contrast and low SNR present in the micrographs. Additionally, the problem is complicated by having to distinguish between particles and damaged particles, contaminated particles, partly formed particles, etc., which, in general we will denote as ‘errors’. We use non-linear, binary support vector machines (SVMs) (Boser *et al.*, 1992; Cortes and Vapnik, 1995) as classifiers due to their good performance in other classification problems [it is particularly used successfully in APP by Arbeláez *et al.* (2011); Zhao *et al.* (2013)], their robustness to noise and their speed. The general idea of the SVM classifier is to find an optimum hyperplane in an n-dimensional space by which two different classes are distinguishable. LIBSVM (Chang and Lin, 2011) is an efficient and widely used implementation of the SVM. This package suggests a variety of kernels to perform the non-linear classification. We use this package with a radial basis function (RBF) kernel to gain a high accuracy in our classification (see Supplementary material for an introduction to SVM).

We use two SVM classifiers. The first classifier is responsible to discriminate between particle objects and any other kind of objects (non-particles and errors). Because of the similarity between errors and particles, the output of the first classifier is not so accurate, and some errors are labelled as particles. Therefore, to reduce the false positives to a feasible extent, the second classifier is dedicated to just focus on distinguishing particles from errors. This strategy was already used by Sorzano *et al.* (2009). Figure 6 shows the behaviour of the classifier for three types of objects. As can be seen, the first classifier passes the error, but the

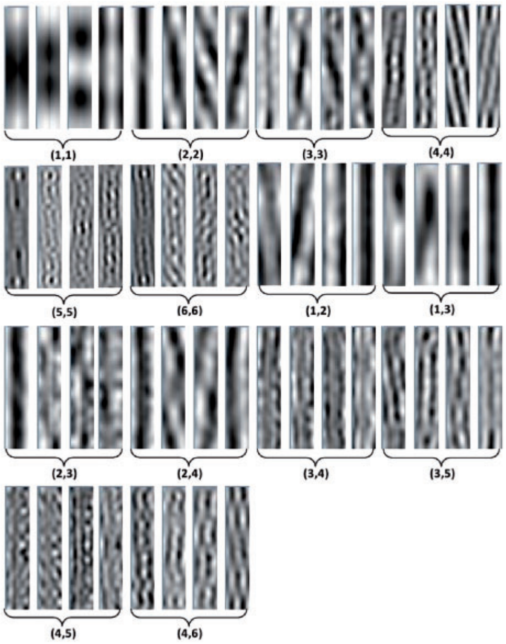

Fig. 4. First four eigenvectors corresponding to the PCA of the cross-correlation functions between two sub-band images for manually picked particles of the KLH dataset. The index under each group of four images shows the contributed sub-band images (*kk*′, in the text)

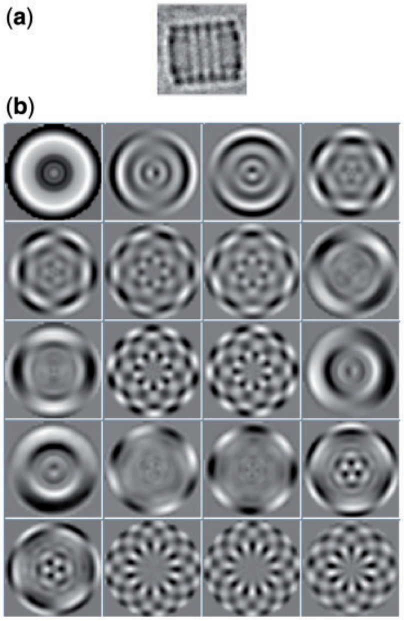

Fig. 5. Eigenvectors (Ponce and Singer, 2011) corresponding to a given template, (**a**) Given particle template. (**b**) First 20 eigenvectors of the template that generate a rotational invariant subspace

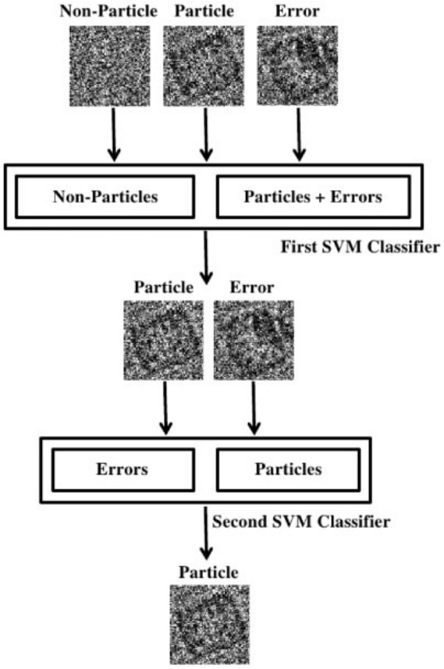

Fig. 6. Behaviour of the classifiers for three types of objects. The first classifier may classify errors as particles, but the second classifier is designed to remove errors from the final result

second classifier rejects it. In fact, the learning process follows the same steps as in the algorithm by Sorzano *et al.* (2009). First, the user picks all particles from the first micrograph and they are clustered in order to construct particle templates. All non-selected locations are assumed to correspond to non-particles and some of them will be later used to train the classifiers. This manual step can be continued with more micrographs to expand the training set. It is worth pointing out that none of the classifiers are trained at this point.

Once the training set is large enough (empirically, at least 30 training particles are required), the process enters a supervised phase. The rotationally invariant subspace as well as the PCAs for the polar, sub-band cross-correlations are calculated on the training particles. Then, feature vectors are calculated for the manually selected particles and non-particles. Finally, the first classifier is trained using this data. At this point, the algorithm tries to automatically pick the next micrograph in the list of micrographs that was not previously manually picked. After suggesting possible particle locations, the user can correct the results by adding those missed particles (false negatives) and removing wrongly picked particles (false positives). After being corrected by the user, the first classifier is retrained using all previous information plus the new set of false negatives and false positives. The second classifier is trained to distinguish between all particles known so far and the set of false positives. This process can be repeated several times on more micrographs till the performance of the classifiers is not further improved by user corrections. This ongoing learning process is particularly interesting because the classifiers carefully adapt to the user's preferences.

When the user is satisfied with the performance of the classifier during the semi-supervised phase, he can go to a fully automatic particle picking mode, in which all micrographs that have not been picked yet are automatically picked (in parallel). At the end of this process, the user can supervise the result and eliminate wrongly picked particles or add missed particles.

3 RESULTS

We applied our APP method on three datasets to assess the speed and accuracy of the proposed algorithm for micrographs with different contrast, density and particle shape and size. These datasets are: KLH (keyhole limpet haemocyanin), adenovirus and helicase. All the experiments were done with the same fixed parameters ($N_b = 4$, $N_h = 6$, $N_{rb} = 20$ and $N_i = 9$) on a single core of a CPU Intel Core i5, 64 bits, 2.53 GHz (of a standard laptop with 4 gigabytes of RAM).

To show the results in a quantitative way, we have used the performance metrics introduced by Langlois and Frank (2011). If TP is the number of true positives, FP the number of false positives and FN the number of false negatives, then these metrics are defined as

- Precision $= \frac{\text{TP}}{\text{FP+TP}}$
- Recall $= \frac{\text{TP}}{\text{FN+TP}}$
- F-measure $= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision+recall}}$

Precision shows the fraction of picked particles by the algorithm that is real particles, and recall indicates the fraction of true particles that are picked by the algorithm. F-measure is a harmonic mean to summarize both recall and precision.

3.1 KLH dataset

This dataset was produced as a general benchmark for APP by Zhu *et al*. (2003). It includes 82 micrographs of KLH particles. A Phillips CM200 TEM was used to record the micrographs on a 2 K × 2 K CCD Tietz camera at a magnification of 66 000× and a voltage of 120 kV. The sampling rate at this magnification was 2.2 Å/pixel.

In the 3DEM Benchmark site (http://i2pc.cnb.csic.es/3dembenchmark) the dataset has been split into two datasets: one with 30 micrographs for training and another one with 50 micrographs for testing. We set the size of particles to 200 pixels, and processed 10% of the local maxima of the correlation map.

To evaluate the connection between the number of particles and accuracy metrics, we calculated precision, recall and F-measure for the test dataset after training the classifier with different numbers of true particles of the training dataset. Figure 7 shows how these three parameters change as the number of training particles is increased. As can be seen in this figure, precision, recall and F-measure of the algorithm are increased if we keep training the classifier with more and more particles. Training with more than 173 particles (13 micrographs), a precision range of [80.0, 85.2] and a recall range of [73.1, 77.8] are achievable. F-measure ranges as a summary of precision and recall in the interval [77.53, 80.06]. Since there is a trade-off between recall and precision, F-measure looks smooth, and changes in a small range.

Training the algorithm with 39 manually picked particles took 3.5 s. After training, the algorithm needs 1 s to suggest new particles on each new micrograph and 2.5 s to retrain after being corrected by the user. Picking particles from 50 micrographs of the test dataset took 51 s, without any parallelization (however, current Xmipp implementation can benefit from multiple CPUs by concurrently picking different micrographs).

According to the 3DEM benchmark, our previous APP method was capable of selecting particles with an average time of 47 s, precision rate 80.94% and recall rate 68.59%. For the new algorithm, the precision and recall rates are reported as 85.16% and 74.81%, respectively, and the average processing time for each micrograph is 1 s. Therefore, the proposed algorithm is very fast and produces accurate results compared with our previous algorithm (Sorzano *et al*., 2009) as well as with the reported results in the APP challenge (Zhu *et al*., 2004).

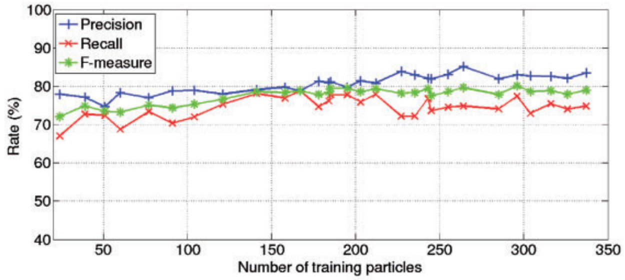

Fig. 7. Three accuracy metrics for the automatic selection of particles from the KLH test dataset according to the number of particles used to train. The vertical axis shows the values for precision (blue), recall (red) and F-measure and the horizontal axis shows the number of training particles

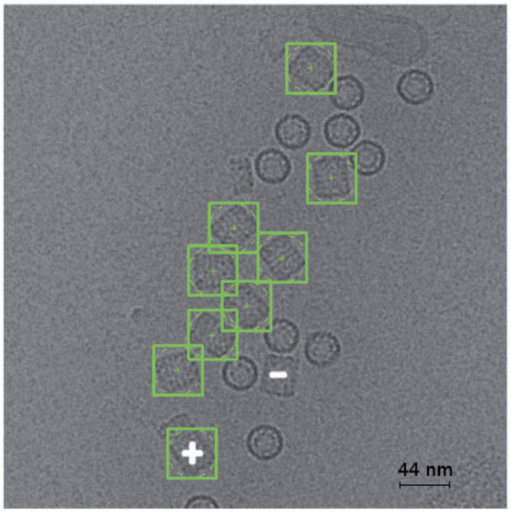

Fig. 8. The result of the algorithm for one micrograph of the KLH dataset with nine particles. Nine objects are boxed in green as particles. There is one false positive and one false negative, which are identified by + and −, respectively

Figure 8 shows the result of the algorithm for one micrograph of the test dataset with nine particles. In this figure, nine green squares show the selected objects, from which one long particle with '+' symbol is a false positive. One particle, marked with '−' symbol, is missed from the final result.

3.2 Adenovirus dataset

In this experiment we examined the reliability of our method by means of a set of micrographs (Pérez-Berná *et al*., 2009) that has lower contrast but higher density and a larger particle size than the KLH dataset (Fig. 9).

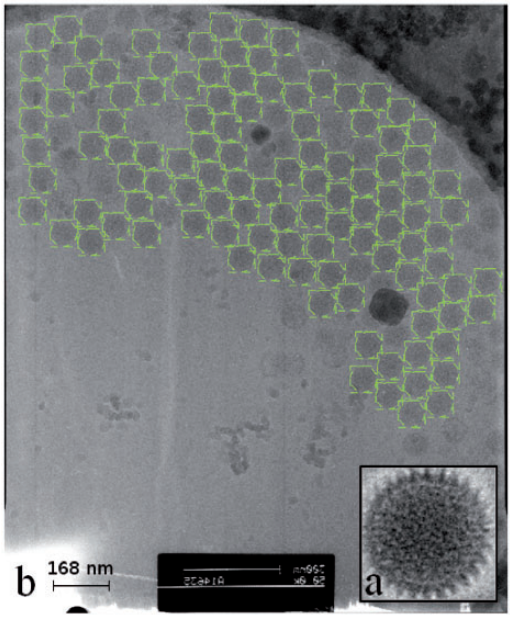

Fig. 9. Automatic selection for one micrograph of the adenovirus dataset, (**a**) The template for cross-correlation. (**b**) The micrograph with 115 automatically selected particles

Adenovirus type 2 is an icosahedral mammalian virus with a molecular mass of 150 MDa, and consists of genomic DNA and structural proteins. Samples were vitrified in liquid ethane, and a FEI Tecnai G2 FEG microscope at working voltage of 200 kV was used to analyse the samples. Micrographs were recorded on a film at a magnification of 5000×, and digitized in a Zeiss Photoscan TD scanner using a step size of 7 µm, which provided a sampling rate of 1.4 Å (particle size at this sampling rate is 600 pixels). True particles for this dataset were manually identified by Pérez-Berná *et al.* (2009).

Like in the previous experiment, the original 305 micrographs are divided into a training dataset with 30 micrographs and a test dataset with 275 micrographs (these two datasets can be obtained from http://i2pc.cnb.csic.es/3dembenchmark/). The sizes of the micrographs are not the same, but we know that $4000 < height < 15\,000$ and $4000 < width < 12\,000$ pixels. The dimensions of the micrographs were internally reduced by a factor of 4. We examined 90% of the local correlation maxima, and assumed that one template was enough for the cross-correlation.

We used 334 particles and 690 non-particles from the first 14 micrographs to fully train the classifier. The classifier was used to automatically pick the particles of the test dataset. This experiment resulted in picking 9216 particles with precision rate 92.17% and recall rate 90.24%. Figure 9b shows the result of the automatic picking for one micrograph of this dataset. As it can be seen in this figure, the algorithm picks 115 particles from this micrograph and particles close to or inside the carbon parts are rejected.

The average processing time per micrograph was 34 s, which is more than for the KLH dataset because it is more dense (90% of the local maxima are checked) and the size of the micrographs and particles were twice and three times larger than the KLH dataset ones, respectively.

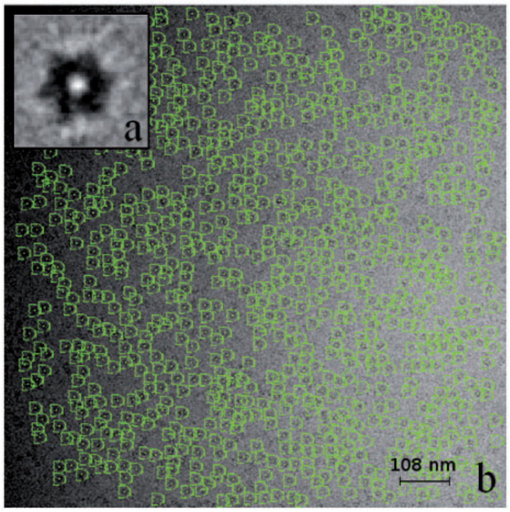

Fig. 10. Automatic selection for the third micrograph of the helicase dataset, (**a**) The template for cross-correlation; (**b**) 874 automatically picked particles from the micrograph

3.3 Helicase dataset

In this section we work with the complex of a helicase and its loading factor with a total molecular mass of 0.39 MDa (V.A. *et al.,* unpublished data). This dataset consists of 160 micrographs of size 4046 × 4046. The CCD of a JEOL JEM-2200FS microscope with magnification 50 000× and voltage 200 kV provided digital micrographs with pixel size 2.16 Å.

The rather dense micrographs of this dataset (the average number of particles in each micrograph is 572) (Fig. 10b) as well as the small particle size make them particularly difficult. Additionally, contrast is especially low due to the cryo-EM conditions. This in-house dataset is not published yet, and no benchmark is available in order to check the accuracy quantitatively, therefore the result is given qualitatively.

We set the size of particles to 100 pixels, and processed 90% of the local maxima to ensure that no particle was missed. To train the classifier, 1037 positive and 2510 negative samples were extracted from the first two micrographs, and then for the third micrograph the algorithm automatically picked 874 particles in 50 s. Figure 10 shows the result for the third micrograph of this dataset, qualitatively in agreement with user expectations.

4 DISCUSSION AND CONCLUSION

In this article, we proposed an APP algorithm that identifies particles from electron micrographs more accurately and faster compared to our previous method by Sorzano *et al.* (2009), already one of the best performing methods. A set of robust shape-related and statistical features are extracted from particle candidate locations (that are distinguished by cross-correlation between the templates and micrograph) and a two-stage classifier decides if each feature vector is a particle or not. Like in

Sorzano *et al*. (2009), the learning process of the classifier is continuous to grab the features of the user desired particles, and the second stage of the classifier is used to concentrate just on making distinctions between particles and errors, which are very similar to particles. Experimental results for three datasets show that this algorithm is able to select particles accurately even from low-contrast and highly dense micrographs.

The feature vector includes three types of features. This feature vector is robust to noise and has a high discrimination power, so that the classifier can distinguish between particles and non-particles. Having two types of features for shape description at the same time helps us to reduce the probability of producing partially similar feature vectors for particles and non-particles. For instance, in case of the KLH dataset, the descriptors at different frequencies are not adequate to properly distinguish between side and top views of the particle. On the other hand, the calculated shape descriptors from rotational invariant subspace are not sufficient to separate errors very similar to the side views of the particle. As a summary, these two types of features conspire efficiently to decrease the false positive rate. Statistical features are also important to catch the properties of the intensity distribution, and prevent the APP from picking particles from the carbon parts of micrographs. The role of this type of feature is clearer for helicase and adenovirus datasets, where more particles lie in dark areas of the micrographs. In Supplementary material, the classification power of the individual features is assessed in depth.

To decrease false positives as much as possible, a two stage SVM classifier is used to classify the feature vectors: the first stage to separate particles and non-particles, and the second stage to remove errors (very similar to particles) from the output of the previous stage. Each SVM component of the classifier is capable of performing the separation with an accuracy of >90% (see Supplementary material). The second component of the classifier plays an important role in reducing the false positive rate (e.g. 15% of wrongly selected particles were discarded by the second classifier in the KLH dataset).

There are six parameters that can be set by the user: number of sub-bands, PCA basis, rotational PCA basis, consecutive sub-bands to be correlated, templates and percentage of local maxima in correlation map to keep. Although these parameters help the user to achieve the highest possible accuracy, they can result in complexity. To moderate this complexity, the value of the first four parameters is set by default, so that the user can focus on adjusting the last two parameters (number of templates and percentage of local maxima to keep).

Regarding the particularities of the datasets, KLH is a dataset with highly contrasted particles, but there are two sources of errors that make it challenging. First, a few background objects and also noisy top views that present features similar to the ones of the particles. Second, KLH can polymerize to some degree. The second classifier is in charge of removing these polymerized particles, and keeping, at the same time, the recall rate at a reasonable level.

For the adenovirus dataset, our algorithm achieved a better accuracy than for the KLH dataset. The reason is that, although the micrographs are denser and have samples with lower contrast, the similarity between particles and non-particles is not as high as in the KLH dataset. The second classifier rejects just 3% of the outputs of the first one. Difficult cases in this dataset are those particles located on carbon areas, which are efficiently distinguished by statistical features.

The speed of the algorithm was examined for the three datasets. According to the density of micrographs, it can go from 1 s to 50 s. Most computations are related to feature extraction, especially the shape descriptors. In order to eliminate this bottleneck and improve the speed even more, in our implementation particle candidates are divided between different threads. In addition to this, the automatic selection of particles from micrographs is performed in parallel. In this way, our APP is extremely fast and can be executed in a very short time.

This algorithm is included in Xmipp 3.0 and downloadable from http://xmipp.cnb.csic.es. The APP is accessible through the protocols described by Scheres *et al*. (2008).

Funding: Economical support from the Spanish Ministry of Economy and Competitiveness (through Grants AIC-A-2011-0638, BFU2009-09331, BIO2010-16566 and CAM(S2010/BMD-2305); NSF (Grant 1114901); postdoctoral Juan de la Cierva Grants (references JCI-2011-10185 and JCI-2010-07594); la Caixa scholarship (to V.A); Formación de Personal Investigador (FPI) fellowship (to A.Z.-P.); CSIC grant (JAE-Predoc to J.M.de la R.-T.); Ramón y Cajal fellowship, (to C.O.S.S.) Y.S. was partially supported by Israel Science Foundation grant 485/10. This work was funded by Instruct, part of the European Strategy Forum on Research Infrastructures (ESFRI) and supported by national member subscriptions.

Conflict of Interest: none declared.

REFERENCES

Arbeláez,P. *et al*. (2011) Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. *J. Struct. Biol*., **175**, 319–328.

Boser,B.E. *et al*. (1992) A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational learning theory, COLT '92*. ACM, New York, NY, USA, pp. 144–152.

Chang,C.-C. and Lin,C.-J. (2011) Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol*., **2**, 27:1–27:27.

Chen,J.Z. and Grigorieff,N. (2007) Signature: a single-particle selection system for molecular electron microscopy. *J. Struct. Biol*., **157**, 168–173.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach Learning*, **20**, 273–297.

Dube,P. *et al*. (1993) The portal protein of bacteriophage SPP1: A DNA pump with 13-fold symmetry. *EMBO J*., **12**, 1303–1309.

Glaeser,R.M. (1971) Limitations to significant information in biological electron microscopy as a result of radiation damage. *J. Ultrastruct. Res*., **36**, 466–482.

Hall,R.J. and Patwardhan,A. (2004) A two step approach for semi-automated particle selection from low contrast cryo-electron micrographs. *J. Struct. Biol*., **145**, 19–28.

Henderson,R. (1995) The potential and limitations of neutrons, electrons and x-rays for atomic resolution microscopy of unstained biological molecules. *J. Mol.r Biol*., **247**, 726–738.

Huang,Z. and Penczek,P.A. (2004) Application of template matching technique to particle detection in electron micrographs. *J. Struct. Biol*., **145**, 29–40.

Langlois,R. and Frank,J. (2011) A clarification of the terms used in comparing semi-automated particle selection algorithms in cryo-em. *J. Struct. Biol*., **175**, 348–352.

Langlois,R. *et al*. (2011) Reference-free particle selection enhanced with semi-supervised machine learning for cryo-electron microscopy. *J. Struct. Biol*., **175**, 353–361.

Ludtke,S.J. *et al.* (1999) EMAN: Semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.*, **128**, 82–97.

Mallick,S.P. *et al.* (2004) Detecting particles in cryo-em micrographs using learned features. *J. Struct. Biol.*, **145**, 52–62.

Nicholson,W. and Glaeser,R. (2001) Review: Automatic particle detection in electron microscopy. *J. Struct. Biol.*, **133**, 90–101.

Ogura,T. and Sato,C. (2004) 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.

Pearson,K. (1901) On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, **2**, 559–572.

Plaisier,J.R. *et al.* (2004) TYSON: Robust searching, sorting, and selecting of single particles in electron micrographs. *J. Struct. Biol.*, **145**, 76–83.

Ponce,C. and Singer,A. (2011) Computing steerable principal components of a large set of images and their rotations. *IEEE T. Image Process.*, **20**, 3051–3062.

Pérez-Berná,A.J. *et al.* (2009) Structure and uncoating of immature adenovirus. *J. Mol. Biol.*, **392**, 547–557.

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

Schatz,M. and van Heel,M. (1990) Invariant classification of molecular views in electron micrographs. *Ultramicroscopy*, **32**, 255–264.

Scheres,S.H.W. *et al.* (2008) Image processing for electron microscopy single-particle analysis using xmipp. *Nat. Protoc.*, **3**, 977–990.

Sigworth,F.J. (2004) Classical detection theory and the cryo-em particle selection problem. *J. Struct. Biol.*, **145**, 111–122.

Sorzano,C.O.S. *et al.* (2004) Normalizing projection images: A study of image normalizing procedures for single particle three-dimensional electron microscopy. *Ultramicroscopy*, **101**, 129–138.

Sorzano,C.O.S. *et al.* (2009) Automatic particle selection from electron micrographs using machine learning techniques. *J. Struct. Biol.*, **167**, 252–260.

Sorzano,C.O.S. *et al.* (2010) A clustering approach to multireference alignment of single-particle projections in electron microscopy. *J. Struct. Biol.*, **171**, 197–206.

Sorzano,C.O.S. *et al.* (2012) Semiautomatic, high-throughput, high-resolution protocol for three-dimensional reconstruction of Single Particles in Electron Microscopy. In: *Nanoimaging: Methods and Protocols. Methods in Molecular Biology*. Humana Press Inc., New York, NY, pp. 171–193.

Volkmann,N. (2004) An approach to automated particle picking from electron micrographs based on reduced representation templates. *J. Struct. Biol.*, **145**, 152–156.

Wong,H.C. *et al.* (2004) Model-based particle picking for cryo-electron microscopy. *J. Struct. Biol.*, **145**, 157–167.

Yu,Z. and Bajaj,C. (2004) Detecting circular and rectangular particles based on geometric feature detection in electron micrographs. *J. Struct. Biol.*, **145**, 168–180.

Zhao,J. *et al.* (2013) Tmacs: A hybrid template matching and classification system for partially-automated particle selection. *J. Struct. Biol.*, **181**, 234–242.

Zhu,Y. *et al.* (2003) Automatic particle detection through efficient hough transforms. *IEEE T. Med. Imaging*, **22**, 1053–1062.

Zhu,Y. *et al.* (2004) Automatic particle selection: results of a comparative study. *J. Struct. Biol.*, **145**, 3–14.