Journal of Structural Biology 183 (2013) 342–353

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques

J. Vargas [a,*], V. Abrishami [a], R. Marabini [b], J.M. de la Rosa-Trevín [a], A. Zaldivar [a], J.M. Carazo [a], C.O.S. Sorzano [a]

[a] Biocomputing Unit, Centro Nacional de Biotecnología-CSIC, C/Darwin 3, 28049 Cantoblanco (Madrid), Spain
[b] Escuela Politécnica Superior, Universidad Autónoma de Madrid, C/Francisco Tomás y Valiente, 28049 Cantoblanco (Madrid), Spain

ARTICLE INFO

Article history:
Received 7 May 2013
Received in revised form 10 July 2013
Accepted 31 July 2013
Available online 6 August 2013

Keywords:
Electron microscopy
Particle picking
Machine learning
Single particle analysis

ABSTRACT

Three-dimensional reconstruction of biological specimens using electron microscopy by single particle methodologies requires the identification and extraction of the imaged particles from the acquired micrographs. Automatic and semiautomatic particle selection approaches can localize these particles, minimizing the user interaction, but at the cost of selecting a non-negligible number of incorrect particles, which can corrupt the final three-dimensional reconstruction. In this work, we present a novel particle quality assessment and sorting method that can separate most erroneously picked particles from correct ones. The proposed method is based on multivariate statistical analysis of a particle set that has been picked previously using any automatic or manual approach. The new method uses different sets of particle descriptors, which are morphology-based, histogram-based and signal to noise analysis based. We have tested our proposed algorithm with experimental data obtaining very satisfactory results. The algorithm is freely available as a part of the Xmipp 3.0 package [http://xmipp.cnb.csic.es].

1. Introduction

Single particle analysis (SPA) techniques based on transmission electron microscopy (TEM) can obtain three-dimensional (3D) reconstructions of biological complexes near atomic resolution (Zhang and Zhou, 2011). However, this high resolution studies require acquiring tens of thousands of projection images. The typical particle selection or picking approach consists in locating the two-dimensional (2D) projections of the biological structure under study within the captured electron micrographs. This process may be cumbersome, laborious and time consuming, especially if done manually, and represents a major bottleneck for SPA of large datasets. However, the success of the reconstruction crucially depends on the number and the quality of the 2D picked particles. In order to develop high-throughput methods that minimize the user iteration in the different processing steps, a number of different automatic and semiautomatic particle picking approaches have been proposed. Automatic particle picking techniques (Chen and Grigorieff, 2007; Adiga et al., 2004; Huang and Penczek, 2004; Kumar et al., 2004; Ogura and Sato, 2005; Plaisier et al., 2004; Rath and Frank, 2004; Roseman, 2004; Singh et al., 2004; Wong et al., 2004) consist in image processing algorithms capable to detect and boxing out the particle projections without the need of any user interaction. These methods are usually fast and provide a large number of particles; however, they may have accuracy and robustness problems, providing a relatively large set of incorrect and erroneously picked particles (false positives). These false positives typically range, depending on the picking algorithm, from fractions of 10% to more than 25% (Zhu et al., 2004). Therefore, after the picking process, it is always required to perform a subsequent manual curation (screening) approach to reject false positives. In turn, semiautomatic picking approaches require the user to provide an initial set of particles, which are manually picked. From this training set, the algorithms learn the kind of objects to be detected and boxed from the micrographs (Arbeláez et al., 2011; Hall and Patwardhan, 2004; Mallick et al., 2004; Ogura and Sato, 2004; Plaisier et al., 2004; Short, 2004; Sorzano et al., 2009; Volkmann, 2004). These methods are halfway between manual and automatic methodologies but they also require a posterior manual particle screening process.

Incorrectly picked particles typically appear because the sample presents some degree of heterogeneity and/or the existence of overlapping or degraded particles in the dataset. Additionally, some of the picked particles usually are strongly affected by noise or may even contain only noise. Finally, the presence of image artifacts such as ice, dust and contaminations, can corrupt the detected particles. In all of these cases, these picked particles represent false positives and must be discarded. Fig. 1 shows an example of typical cases of correctly and incorrectly picked particles. In Fig. 1(a), we see a correctly picked particle (true positive) of *Bovine papillomavirus* (Wolf et al., 2010). In turn, Fig. 1(b), (c) and (d) show examples

* Corresponding author. Fax: +34 585 4506.
E-mail address: jvargas@cnb.csic.es (J. Vargas).

1047-8477/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jsb.2013.07.015

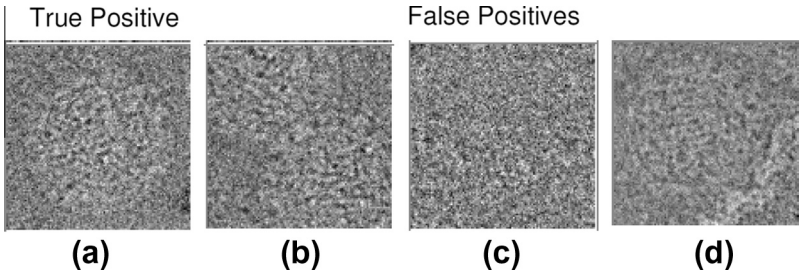

Fig.1. Example of three kinds of typical automatic or semiautomatic particle picking problems; (a) correctly picked particle (true positive) of *Bovine papillomavirus*, two overlapping particles (b), only noise image (c) and particle affected by an artifact (d).

of two overlapping particles, an only noise image and a particle affected by an artifact, respectively.

Post-picking methodologies, based on processing the output of automatic and/or semiautomatic particle picking methods, have been previously proposed to separate particle images from non-particle ones (Norousi et al., 2013; Sorzano et al., 2004). Since these two strategies are conceptually close to our newly proposed approach, we will describe them in some depth in the following. In (Norousi et al., 2013), a supervised discriminative post-picking approach based on characteristic features calculated from the boxed out images is presented. This method requires the user to provide a training dataset from the previously picked particles. The classification is based on the extraction of distinct features, such as: radially weighted average intensity, phase symmetry and dark dot dispersion. First, the radially weighted average intensity is calculated as a weighted sum of the pixel intensities. The weighting is inversely proportional to the pixel's Euclidean distance from the image center. Note that this descriptor does not obtain reliable values in cases where the particle centroid is not placed exactly at the image center. In these cases, the particles will be incorrectly considered as false positives. Additionally, the phase symmetry is computed from a set of 2D wavelets that extracts local frequency information (Kovesi, 1997). In (Norousi et al., 2013), it is claimed that non-particle images will have larger locally symmetric areas, but this consideration may not be general. Finally, the dark dot dispersion descriptor consists in convolving the image with a 2D symmetric Gaussian kernel and binarizing the resultant image using the 0.95 intensity quartile. The authors establish that the average distance between the dark detected regions is larger in particle images than in non-particle images. Observe that the good performance of this descriptor depends strongly on the signal-to-noise ratio (SNR) of the images. Additionally, in (Sorzano et al., 2004) it is presented a post-picking particle sorting method to determine the quality of the input boxed images and identifying erroneous particles. This method uses as descriptors both the radial average intensity and the image histogram. Note that as mentioned before, the radial average descriptor presents problems in cases where the particle centroid is not placed in the image center.

Building on the large body of experience in the field, which has been briefly presented in previous paragraphs, we present in this work a new approach to particle quality sorting that outperforms previous approaches while being computationally efficient. The main objective of this work is to detect outliers (incorrectly picked particles) from a previously picked dataset and not to be very accurate in the fine assessment of correctly picked particles. The input of the algorithm is a previously picked particle dataset that can be affected by outliers and may be coming from any manual, semiautomatic or automatic particle picking method. The unique requirement of our new approach is that there is a majority of correctly picked particles in the dataset. For each of the provided particles, several different types of descriptors are obtained, that are morphology, histogram and noise-based. Morphology descriptors encode information about the shape of the particles. Histogram descriptors give statistical intensity information of the particle images. Finally, noise-related descriptors allow for the separation of noise-only images from those containing signal and noise. For each particle and type of descriptor we compose a vector and, consequently, we will have three vectors per particle. Stacking the vectors of the same descriptor class and of all particles, we compose three descriptor matrices. Using a principal component analysis (PCA) dimensionality reduction approach (Roweis, 1998), we obtain an error score (z-score) for each particle taking into account the descriptor matrices. Furthermore, we study the statistical distribution of our proposed z-score under a set of simple hypothesis, reaching the conclusion that z-score values around 3 should be appropriated, specially when performing the sorting in a fully automatic, parameter-less, high throughput way. Low values of this z-score mean high reliability of the particle under study. Note that a reliable z-score measure is of high importance as it can be used to discriminate between true positive and false positive particles. Additionally, this reliable particle measure can also be utilized as a weighting parameter of the different projections in further processing steps, as for example the three-dimensional density reconstruction, although we have not exploited this issue in the present work.

The paper is organized as follows. In next section, we present the particle quality assessment and sorting method. In Section 3, we show some experimental data results and, finally, the discussion and conclusions are given.

2. Methods

In this section, we present the proposed approach to obtain the morphology, noise-based and histogram descriptors, as well as, the method to compute the particle score from them.

2.1. Morphology descriptors

The morphology descriptors obtain image features that enable discrimination of incorrect particlesbased on their general morphology/shape while eliminating overlapping particles. A good example of the kind of incorrectly picked particles that these morphology descriptors will remove can be seen in Fig. 1. Note that the shape of the particles shown in Fig. 1(b) and (d) is very different from the particle shape shown in Fig. 1(a). Therefore, in these cases, the morphology descriptors will provide valuable information to differentiate between these correctly and incorrectly selected particles. In this work, we use two different sets of morphology descriptors. The first one is based on an image normalization process using the spiral phase transform (SPT) method (Larkin et al., 2001) and Fourier filtering. The second one is derived from the particle autocorrelation map, which is sensitive to the particle shape, at the same time that it is shift invariant.

The intensity of any image can be arbitrarily modeled without lost of generality as

$$I(x,y) = A(x,y) + B(x,y)\cos[\Phi(x,y)] + \eta(x,y) \tag{1}$$

where $A = A(x,y)$ is the background signal, $B = B(x,y)$ is the contrast or amplitude term, $\eta = \eta(x,y)$ is the additive noise and $\Phi = \Phi(x,y)$ is a phase map that encodes the shape or morphology of the imaged object. In SPA, the size of the boxed particles is usually small and, therefore, we can assume the background signal to be approximately constant. From Expression (1) we may define $\tilde{I}$ as

$$\tilde{I} = FT^{-1}[H \cdot FT[I]] \cong \tilde{B}\cos[\Phi] \tag{2}$$

with $\tilde{I} = \tilde{I}(x,y)$ being the normalized and low-pass filtered image, $\tilde{B} = \tilde{B}(x,y)$ the contrast or amplitude signal after the Fourier filtering, $FT[\cdot]$ the Fourier Transform operator and H an isotropic low-pass frequency filter defined as

$$H(R) = \begin{cases} \exp[-R^2/2\sigma^2] & R \neq 0 \\ 0 & R = 0 \end{cases} \tag{3}$$

where $R = \sqrt{\omega_x^2 + \omega_y^2}$, with ω_x and ω_y the normalized frequency components in the Fourier space and σ corresponding to the filter standard deviation. Note that in our case, we are only interested in the global shape of the particles and not in their high resolution details. Additionally, observe that only one particle is presented in each correctly picked image. Therefore, we can perform a strong low-pass filtering to detect the global shape of the particle with typical values of σ around 0.1 px^{-1}. Finally, note from Eq. (2), that the background suppressed and filtered particle $\tilde{I}$ can be factorized in two terms. On one hand, the contrast or amplitude term, $\tilde{B}$, that can be interpreted as a quality map (Ströbel, 1996). On the other, the phase or shape term, $\cos[\Phi]$, that gives information about the shape of the particle (Kovesi, 1997, 2002). Note that the dynamic range of this term (shape term) is limited to $[-1, 1]$ in arbitrary units (a.u) and, therefore, this map is of great importance to detect and segment the particle shape. In order to clarify this point, in Fig. 2 we show a pattern as given in Eq. (2) and its corresponding $\tilde{B}$ and $\cos[\Phi]$ terms. As can be seen from Fig. 2(c), the phase term gives information about the pattern shape, in this case a simple pattern of white squares on a black background. In order to construct the amplitude and contrast maps, we use the spiral phase transform (Larkin et al., 2001).

The spiral phase transform (SPT) mathematical operator corresponds to (Larkin et al., 2001)

$$SPT[\cdot] = FT^{-1}\left[\left(\frac{\omega_x + i\omega_y}{\sqrt{\omega_x^2 + \omega_y^2}}\right) FT[\cdot]\right] \tag{4}$$

If we apply the SPT operator to the $\tilde{I}$ signal given in Expression (2), we obtain the quasi-quadrature respective signal of $\tilde{I}$ (Larkin et al., 2001), given by

$$SPT[\tilde{I}] = i\exp[iD]\tilde{B}\sin[\Phi] \tag{5}$$

where $i = \sqrt{-1}$ and $D = D(x,y)$ is the direction phase map, that is defined as the angle formed by the phase gradient with respect to the x-axis and can be obtained as

$$D = \arctan\left[\frac{\nabla_y \Phi}{\nabla_x \Phi}\right] \tag{6}$$

Note that $\Phi = \Phi(x,y)$ is an unknown quantity and then, it is not possible to obtain the direction map from Expression (6). Instead of D, we can compute the orientation map as

$$\theta = \arctan\left[\frac{\nabla_y \tilde{I}}{\nabla_x \tilde{I}}\right] = \pm D \tag{7}$$

Observe that θ and D have the same magnitude but possible different local signs. We can rewrite Eq. (5) using Eq. (7) as

$$SPT[\tilde{I}] = i\exp[\mp i\theta]\tilde{B}\sin[\Phi] \tag{8}$$

Note that in Eq. (8) appears a sign ambiguity in θ. We can obtain the phase Φ from Eq. (2) and (8) as

$$\Phi = \pm\arctan\left[\frac{-i\exp[-i\theta]SPT[\tilde{I}]}{\tilde{I}}\right] \tag{9}$$

Observe that Eq. (9) is affected also by the sign ambiguity problem. However, the phase or shape term, that corresponds to the cosine of this phase, is not affected by the sign ambiguity because the cosine is an even mathematical function. The shape term is given, therefore, by

$$\cos[\Phi] = \cos\left[\arctan\left[\frac{-i\exp\left[-i\theta]SPT[\tilde{I}\right]}{\tilde{I}}\right]\right] \tag{10}$$

The modulating signal $\tilde{B}$ can be obtained as

$$\tilde{B} = \sqrt{\left(\cos\left[\arctan\left[\frac{-i\exp[-i\theta]SPT[\tilde{I}]}{\tilde{I}}\right]\right]\right)^2 + \tilde{I}^2} \tag{11}$$

In order to obtain the first set of morphology descriptors, we segment the global shape of the particle, binarizing and labeling the shape term, $\cos[\Phi]$ and capturing the region with largest area. Note that the labeling is a morphological operation for identifying each object in a binary image, which consists on marking each connected components in a 2D binary image with different integer numbers. Additionally, binarizing is performed directly by $\cos[\Phi] > 0$. We fit an ellipse to the obtained binarized region and we compute the morphology descriptors, which are (1) the Euclidean distance between the region centroid and the image center, (2) the major and (3) minor axis of a fitted ellipse and the (4) area of this segmented region. These descriptors encode the information

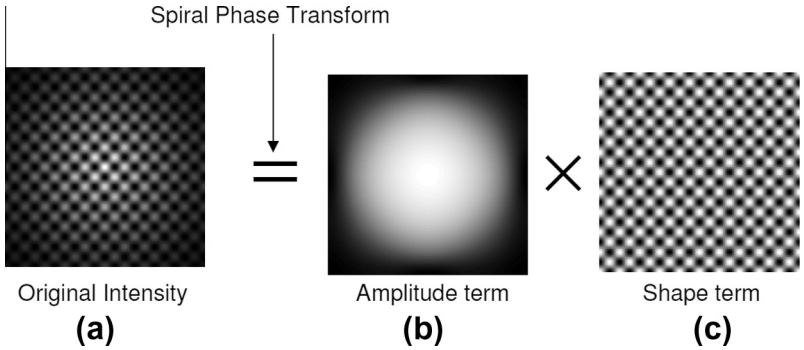

Fig.2. Example pattern (a) and its corresponding amplitude (b) and phase (c) terms.

about the shape of the detected object. Obviously, the shape of a particle that overlaps with another particle, or of a projection of a different biological object, will be very different in general. Therefore, the fitted ellipse, as well as the other morphological descriptors, will be very different also. In Fig. 3, we present an example of this process showing the different obtained maps. Fig. 3(a) shows a previously picked particle and the obtained shape term (b). In Figs. 3(c) and (d) we show the computed binarized map from the shape term and the corresponding labeled map. Finally, Figs. 3(e) and (f) show the captured region with largest area and the obtained morphology descriptors. From Fig. 3(f), we can see that an incorrectly picked particle, with its shape significantly different from the one of a correctly picked one, will present morphology descriptors notably different from those of the latter ones. This case typically appears when two particles overlap, but also occurs when the picking algorithm detect something that it is not a particle, or a projection of a different biological object. For each particle, we compose a 4×1 vector (v_{m1}) formed by the morphology descriptors presented before.

The other set of morphology descriptors used by our procedure are acquired from the autocorrelation map of the different particle images. Note that the autocorrelation depends on the particle shape and, additionally, is shift invariant; therefore this map can be used without problems in cases where the particle centroids are not placed exactly in the image center. We then define this second set of morphological descriptors as the radial average of each particle along a number of equi-spaced radii. This number has typically a value of 100, but for very large or very small particles this number must be modified. In this way, we obtain a 100×1 vector (v_{m2}).

2.2. Noise-based descriptors

Noise-based descriptors obtain image features that enable discriminating between correctly picked particles and very noisy particles or images without particle corresponding to only noise projections. In this work, we propose to use the eigenvalue distribution of the image covariance matrix as a noise-based descriptor. The image of a particle affected by additive noise can be described by

$$I = I_p + I_n \tag{12}$$

where I_p and I_n are the particle image and a random matrix with independent identically distributed entries with zero mean and variance σ, respectively. The covariance matrix is given as

$$C = I^T I = (I_p + I_n)^T (I_p + I_n) \tag{13}$$

Let start first with the case in which we have a very noisy image, then $I_p << I_n$. In this case, we can consider that $C \cong C_n = I_n^T I_n$ and the empirical eigenvalues distribution of $C \cong C_n$ converges to a non-random distribution function when $N \to \infty$, with N the image size, whose density is given by (Müller, 2004)

$$P_C(\lambda) = \begin{cases} \frac{1}{2\pi} \frac{\sqrt{\lambda_{\max} - \lambda}}{\lambda}, & \lambda \in (0,\ \lambda_{\max}) \\ 0 & \textit{elsewhere} \end{cases} \tag{14}$$

with

$$\lambda_{\max} = 4\sigma \tag{15}$$

and the eigenvalues are given by direct integration of $P_C(\lambda)$ as

$$\lambda = \int_0^x \frac{1}{2\pi} \frac{\sqrt{\lambda_{\max} - x}}{x}, \quad x < \lambda_{\max} \tag{16}$$

On the other hand, we can consider the case where the noise is almost negligible. In this situation, we have that $I \cong I_p$ and, therefore, $C \cong C_p = I_p^T I_p$. The eigenvalues distribution of this image covariance matrix is characterized by a small set of large eigenvalues that describes all the information contained in the image and the rest of values equal to zero (Basu et al., 2010; Vargas et al., 2013). Note that for a square image of size $N \times N$, the number of degrees of freedom corresponds to N^2. However, for particle

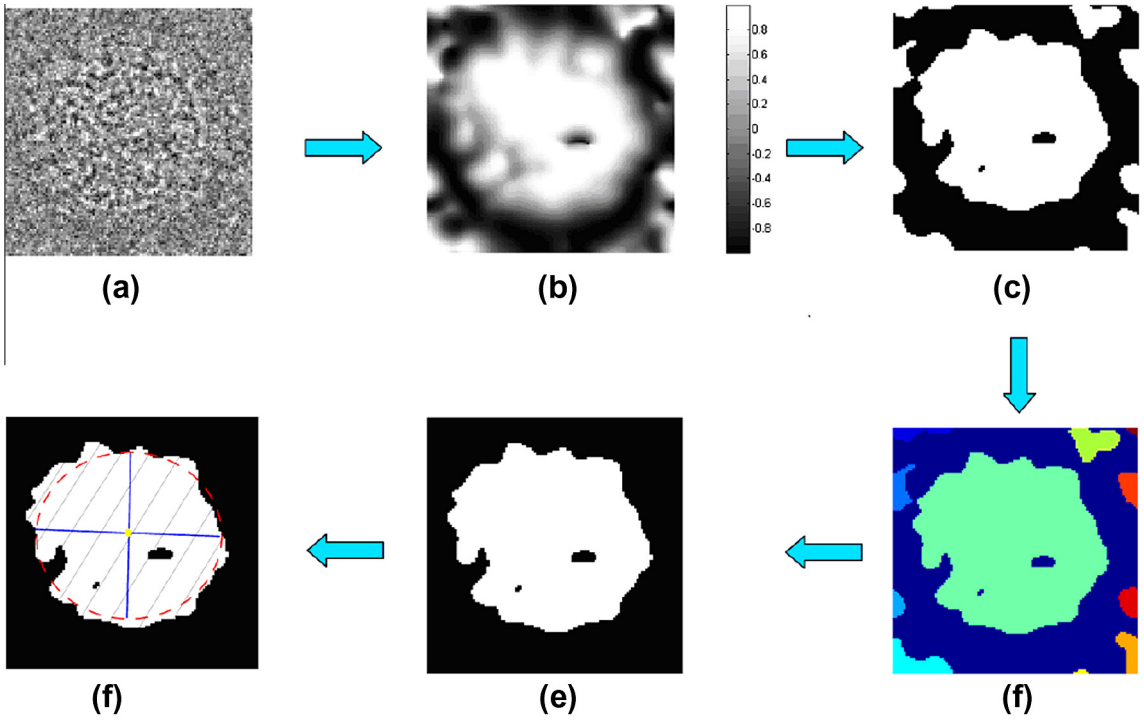

Fig.3. Different processing steps required to compute the morphology descriptors based on the phase term segmentation, where (a) shows a previously picked particle, (b) is the obtained shape term, (c) and (d) show the computed binarized map from the shape term and the corresponding labeled map. Finally, (e) and (f) show the captured region with largest area and the obtained morphology descriptors.

images in particular, and natural pictures in general, the pixel values are not independent and can be correlated far away from their location. Therefore, the number of actual degrees of freedom is strongly reduced and all the information can be coded in a small set of large eigenvalues. Finally, we must address the case where $I_p \cong I_n$. In this situation the particle and noise images have similar power and the eigenvalue distribution is dominated for small values of the eigenvalues by the noise term, while for large values by the particle signal.

The actual eigenvalue distribution of C can be obtained by the singular value decomposition (SVD) factorization algorithm. The SVD algorithm consists in a unique factorization of a matrix of the form,

$$C = UDV^T \tag{17}$$

where C is a real symmetric matrix, U and V^T are orthogonal matrices and D is a diagonal matrix. The entries of D are non-zero only on the diagonal, and are known as the singular values of C. In our case, these values are real and positive valued as C is a symmetric matrix. Additionally, for symmetric matrices, eigenvalues and singular values are identical. Therefore, we can obtain the eigenvalues of C from the diagonal of the D matrix computed by the SVD decomposition. In Fig. 4, we show a particle projection (a), an only-noise image extracted from the same micrograph and with the same noise variance than the particle projection (b), and a plot of the respective eigenvalues distribution, where the solid black line corresponds to the particle projection while the dashed curve refers to the only noise image (c). Observe that at low eigenvalue index, the eigenvalues of the particle projection and those of the only-noise image are very similar. However, the eigenvalues are significantly different for large eigenvalue index. We use this idea to define a noise-related descriptor, and we compose for each previously picked particle a 20×1 vector (v_n) formed by the first twenty largest eigenvalues in magnitude.

2.3. Histogram descriptors

Histogram descriptors obtain image features that enable discrimination of incorrect particles affected by image artifacts caused by ice variations, dust and contaminations, among other factors. To this end, we calculate the histogram of the imaged particles and measure the percentiles—i.e., the values of the intensity below which a certain percent of observations fall—10%, 20%, 30%, 40%, 50%, 60%, 70%, 80 and 90%. With these percentiles, we compose a 9×1 vector (v_h) for each boxed out particle. In Fig. 5, we show a particle projection that can be considered good (a) and a particle projection distorted by artifacts (b). Additionally, we also show the respective image histograms. As can be seen from Fig. 5, the histograms are significantly different.

2.4. Z-Score evaluation

Using the previously introduced descriptors, we may define a z-score index that will be the base of our particle screening algorithm. To obtain the z-score index for every particle, we firstobtain the average feature vectors, $\bar{v}_{m1}$, $\bar{v}_{m2}$, $\bar{v}_n$ and $\bar{v}_h$, and we compute from them the zero mean feature vectors, that for the ith element are given by

$$\begin{aligned}(\tilde{v}_{m1})_i &= (v_{m1})_i - \bar{v}_{m1}\\ (\tilde{v}_{m2})_i &= (v_{m2})_i - \bar{v}_{m2}\\ (\tilde{v}_n)_i &= (v_n)_i - \bar{v}_n\\ (\tilde{v}_h)_i &= (v_h)_i - \bar{v}_h\end{aligned} \tag{18}$$

We stack these zero mean feature vectors into four descriptor matrices named morphology$_1$, morphology$_2$, noise and histogram matrices, which have dimensions of $4 \times Np$, $100 \times Np$ $20 \times Np$ and $9 \times Np$ respectively, with Np the number of initially picked particles. These matrices are given by

$$\begin{aligned}M_{m1} &= [(\tilde{v}_{m1})_1,\ (\tilde{v}_{m1})_2,\ \ldots,\ (\tilde{v}_{m1})_{N_p}]\\ M_{m2} &= [(\tilde{v}_{m2})_1,\ (\tilde{v}_{m2})_2,\ \ldots,\ (\tilde{v}_{m2})_{N_p}]\\ M_n &= [(\tilde{v}_n)_1,\ (\tilde{v}_n)_2,\ \ldots,\ (\tilde{v}_n)_{N_p}]\\ M_h &= [(\tilde{v}_h)_1,\ (\tilde{v}_h)_2,\ \ldots,\ (\tilde{v}_h)_{N_p}]\end{aligned} \tag{19}$$

where $(\cdot)_i$ denotes the ith particle. For each matrix, we obtain a principal component analysis (PCA) basis composed by two vectors. PCA is a widely used dimensionality reduction technique in data analysis (Gonzalez and Woods, 2007) that involves a mathematical procedure, which transforms a number of possibly correlated vectors into a smaller number of uncorrelated ones called the principal components. We have used an efficient algorithm in terms of space and time to retrieve the PCA basis from large collections of data, which is the case we are interested in (Roweis, 1998). We project the different zero mean feature vectors to their corresponding PCA basis and we obtain, therefore, the vectors components in these PCA bases as

$$\begin{aligned}P_{m1} &= [(p_{m1})_1,\ (p_{m1})_2,\ \ldots,\ (p_{m1})_{N_p}]\\ P_{m2} &= [(p_{m2})_1,\ (p_{m2})_2,\ \ldots,\ (p_{m2})_{N_p}]\\ P_n &= [(p_n)_1,\ (p_n)_2,\ \ldots,\ (p_n)_{N_p}]\\ P_h &= [(p_h)_1,\ (p_h)_2,\ \ldots,\ (p_h)_{N_p}]\end{aligned} \tag{20}$$

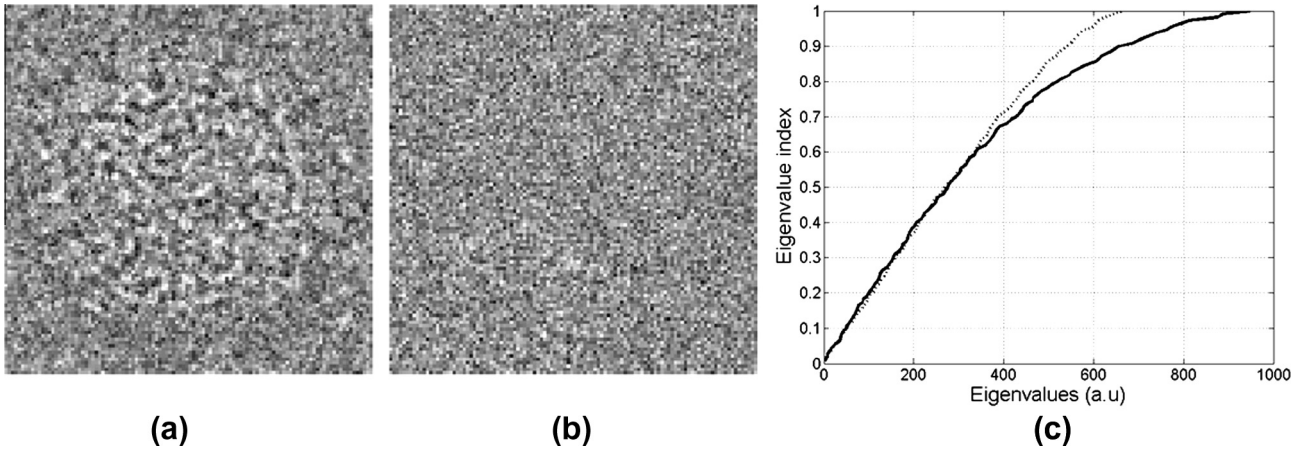

Fig.4. Particle (a), only-noise image (b) and corresponding eigenvalue distribution (c), where the solid black line corresponds to the particle image while the dashed curve refers to the only noise image.

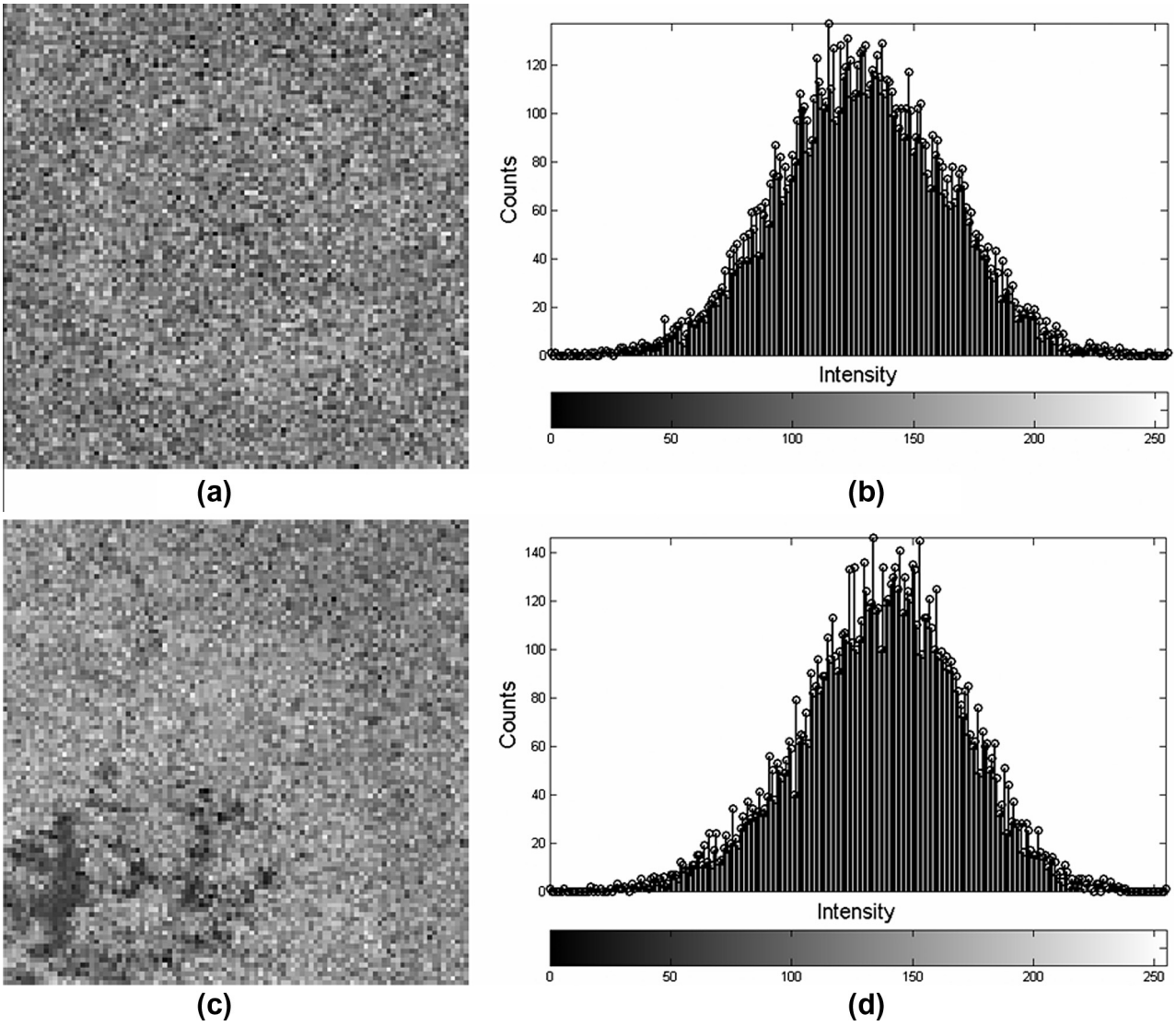

Fig.5. Particle image (a), corresponding intensity histogram distribution (b), particle image affected by an artifact (c) and corresponding intensity histogram distribution (d).

where $(p_{m1})_i$, $(p_{m2})_i$, $(p_n)_i$ and $(p_h)_i$ are the projection of the ith particle zero-mean feature vectors into the PCA bases computed from M_{m1}, M_{m2}, M_n and M_h, respectively. Note that the dimension of the different projection vectors is the number of principal components used, that in our case are two for every case. We can now obtain a measure of similarity for the different particles using the Mahalanobis distance (Filzmoser, 2005). For each of the projection vectors, we compute the respective covariance matrices for the ith particle as, $(C_k)_i = (p_k)_i(p_k)_i^T$, $k = \{m_1, m_2, n, h\}$ and, we estimate the covariance matrix of each of the projection sets as

$$\begin{aligned}
\tilde{C}_{m1} &= (1/N_p)\sum_{i=1}^{N_p}(C_{m1})_i \\
\tilde{C}_{m2} &= (1/N_p)\sum_{i=1}^{N_p}(C_{m2})_i \\
\tilde{C}_{n} &= (1/N_p)\sum_{i=1}^{N_p}(C_{n})_i \\
\tilde{C}_{h} &= (1/N_p)\sum_{i=1}^{N_p}(C_{h})_i
\end{aligned} \tag{21}$$

Finally, the different particle Mahalanobis distances, which are interpreted as an error index or z-score, are given by

$$\begin{aligned}
(Z_{m1})_i &= \sqrt{(p_{m1})_i^T \tilde{C}_{m1}^{-1}(p_{m1})_i} \\
(Z_{m2})_i &= \sqrt{(p_{m2})_i^T \tilde{C}_{m2}^{-1}(p_{m2})_i} \\
(Z_{n})_i &= \sqrt{(p_{n})_i^T \tilde{C}_{n}^{-1}(p_{n})_i} \\
(Z_{h})_i &= \sqrt{(p_{h})_i^T \tilde{C}_{h}^{-1}(p_{h})_i}
\end{aligned} \tag{22}$$

Finally, we can combine these distances in a unique z-score taking their infinity norm as

$$(Z)_i = \|(Z_{m1})_i,\ (Z_{m2})_i,\ (Z_n)_i,\ (Z_h)_i\|_\infty \tag{23}$$

If the original projections (p) are normally distributed, then the square of the Mahalanobis distance approximately follow a χ^2 distribution with degrees of freedom equal to the dimension of the projection vectors (Krzanowski, 2000), which is two in our case. Therefore, we can consider that an outlier has a z-score of $(Z)_i \geqslant 3$ (if the data is normally distributed, then only 1.1% of the population is identified as outlier).

2.5. Particle screening

Based on the z-score index introduced before, we propose a particle screening approach. The particle screening process, that consists of separating the correct particles (true positives) from the false ones (false positives or outliers), can be done in a fully automatic, semi-supervised or in an iterative way.

2.5.1. Automatic

The fully automatic process consists in selecting a cut-off or threshold value and to reject any particle having a z-score larger than this quantity. Usually, it is recommended to select a threshold value about 3 that, as explained before, means that we consider that an outlier has a probability associated with its Mahalanobis distance of 0.01 or less, assuming that the projections are normally distributed. Note that in this mode there is no user iteration during the screening procedure allowing for high-throughput processing. However, the main drawback of this method is that some correctly picked particles may also be rejected (false negatives).

2.5.2. Semi-supervised

In this case, a cut-off or threshold value, typically of 2.5 or lower, is selected and all the particles having a z-score larger than this quantity are first disabled. However, these initially disabled particles are presented to the user in order to be supervised. Therefore, particles that have been erroneously rejected can be enabled again. In order to help the user in this process, we have developed a friendly graphical interface in Xmipp, in which the different particles are sorted according to their z-score value. Therefore, the user only has to evaluate the last particles, with largest z-score, that have been initially disabled. Note that the main advantage of this procedure is that the user only has to screen a very reduced number of particles and not the whole dataset. This limited supervising process makes this task less cumbersome and time consuming and, at the same time, reduces the probability of performing mistakes in the manual screening process. In Fig. 6 we show two different GUIs used in Xmipp to perform the screening process. In Fig. 6(a) we show a gallery of particles used in the semi-supervised particle screening process. In this example, the dataset is composed of 231 previously picked particles and the user only has to verify the validity of the last 24, that are marked in Fig. 6 with a red cross. Additionally, in Fig. 6(b) we show another GUI to perform the screening process where the micrographs with the detected particles are presented. Note that the box around the particles is shown with different colors according to their z-score value. The particles with highest z-scores, which are marked with red color squares, can be disabled directly using this graphical tool.

2.5.3. Iterative

We let the users run an additional processing option that can be used in an automatic or semi-supervised way. In both cases (automatic or semi-supervised) the particle screening process runs iteratively, being the output of eachiteration (cured particles), the input of the next one. This process is repeated until there is no any particle with a z-score larger than the selected cut-off or threshold. This procedure has a statistical advantage since, in each iteration, the covariance matrix needed by PCA is less affected by outliers. Indeed, this approach is at the core of many robust statistical procedures.

3. Results

We have used the data available at the 3DEM Benchmark site (http://i2pc.cnb.csic.es/3dembencmark), that provides a robust computational infrastructure designed to support developers in the benchmarking process of their algorithms. Two different datasets are available in the web site for downloading and processing. One corresponds to the classical Keyhole Limpet Hemocyanin (KLH) particles, while the second one is new, internally acquired, and corresponds to micrographs of Human adenovirus type 2 (Ad2 ts1) samples. The accuracy assessment is performed automatically by the website, comparing the developer reported particles coordinates with respect to ground-truth coordinates, which have been previously picked up manually. This comparison is made in terms of some figures of merit that are precision, recall and *F*-measure rates (Langlois and Frank, 2011) that are given by

$$\begin{aligned} P &= \frac{TP}{TP+FP} \\ R &= \frac{TP}{FN+TP} \\ F &= 2\frac{P\cdot R}{P+R} \end{aligned} \tag{24}$$

where, *TP*, *FN*, *FP*, *P*, *R*, and *F* mean true positive, false negative, false positive, precision rate, recall rate and *F*-measure respectively. The precision rate establishes the probability of detecting true positive particles with respect to the total number of picked particles (true and false) by the particle picking algorithm. The recall rate determines the probability of detecting true particles with respect to the total number of good particles presented in the micrographs. Finally, the *F*-measure is a harmonic mean that accounts for both the precision and recall rates.

3.1. Keyhole Limpet Hemocyanin particles (KLH)

The first set of micrographs is comprised by cryo images of Keyhole Limpet Hemocyanin particles (KLH) recorded on a charge-coupled device (CCD) at 120 kV (Zhu et al., 2003). This dataset was used in the particle selection bakeoff (Zhu et al., 2004) and is composed by approximately 1000 particles that were manually selected by Mouche (one of the participants), composing the reference set. In this work, we have used the same reference set as ground truth to obtain our results (figures of merit). Our proposed particle sorting approach operates on a set of particles previously picked by any automatic or manual procedure. For the sake of simplicity, in this work we have used the output of the Xmipp picking particle approach (Abrishami et al., 2013), which is one of the currently best performing methods. We consider two situations. In the first case, the previous particle picking algorithm (Abrishami et al., 2013) is run under conditions in which the picking of particles is very restrictive, detecting a small number of particles (747) and acquiring a large number of false negatives while a small number of false positives. In the second situation, the picking algorithm obtains a large number of particles (1219), with a small number of false negatives while a large number of false positives. In both cases, the number of particles is too small for any practical structural studies, but this reduced size allows performing detailed comparisons with manual procedures (Zhu et al., 2004). The size of each picked image is always of 250×250 px. The proposed particle sorting approach is now used on these two sets of the KLH data.

3.1.1. First case

In the first case, we have 747 previously picked particles, as explained before. In Fig. 7 we show the obtained figures of merit for different z-scores thresholds, where red, blue and black lines correspond to the precision, recall and *F*-measure rates, respectively. The green lines are the percentage of particles taken into account (not disabled) in the automatic case (Fig. 7(a) and (d)), and the percentage of particles supervised in the semi-supervised case (Fig. 7(c) and (f)). Additionally, in Fig. 7 we show colored bands that correspond to confidence bands. These bands are generated obtaining, on one hand, the different figures of merit from the particle picking algorithm without any particle screening process, and

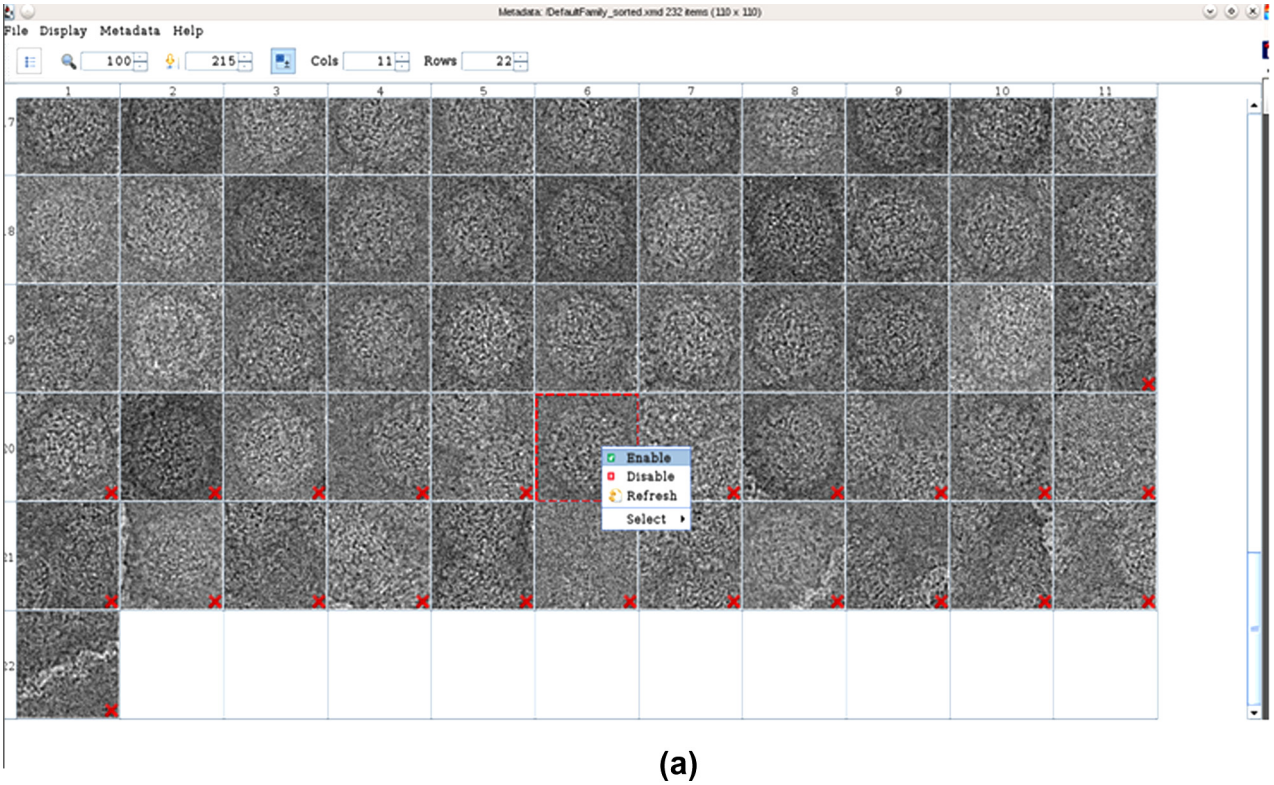

Fig.6. Graphical user interfaces used in Xmipp to perform the screening process using a z-score threshold. Gallery of images with initially discarded particles marked with a red cross (a). Color representation of the z-score assigned to each particle in a processed micrograph (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

on the other hand, supervising manually every particle in the dataset after the particle picking task (fully-supervised). We have represented the precision and recall confident bands with red and blue colors. On the other hand, the *F*-measure confidence band is defined by the region between the two dashed black lines. When we supervise manually every particle (fully-supervised) we obtain precision, recall and *F*-measure rates of 92.56%, 49.77% and 64.73%, respectively. On the other hand, the results obtained from the particle picking task without any screening process correspond to precision, recall and *F*-measure of 84.70%, 51.70% and 64.21%, respectively. Note that these bands are the same in all the plots of Fig. 7.

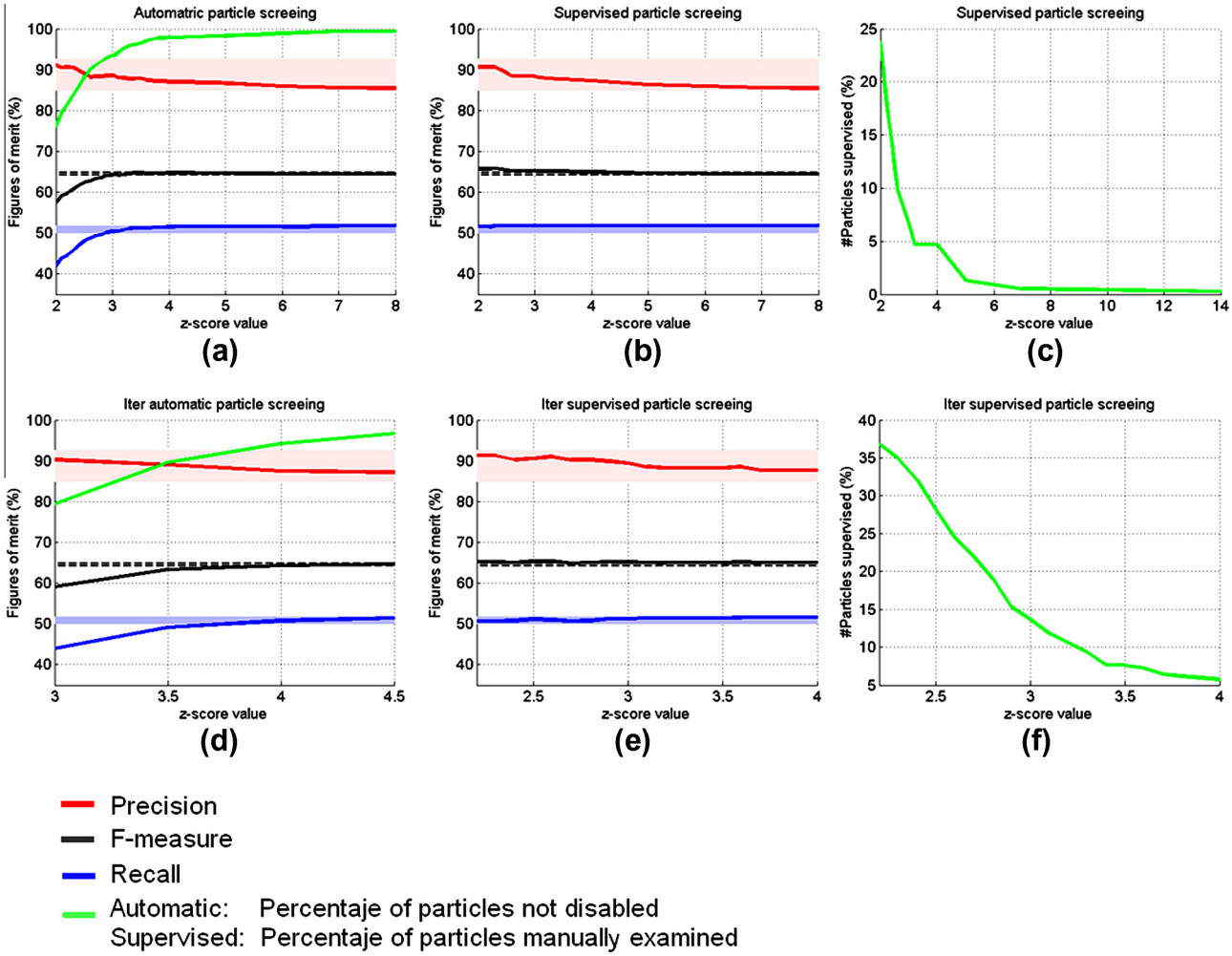

Fig.7. Figures of merit where red, blue and black lines correspond to the precision, recall and *F*-measure rates, respectively, and obtained in the first case of the KLH dataset. The precision and recall confident bands appear with red and blue colors. Additionally, the *F*-measure confidence band is defined by the region between the two dashed black lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In the automatic case (Fig. 7(a)), we see that when a small z-score threshold of 2 is used, we obtain a precision rate close to the fully-supervised case that corresponds to 91.02% and we disable the 23.96% of the particles. Observe that in this situation, the recall value is low (42.07%) compared with the fully-supervised result (49.77%) because a large number of true positive particles are also discarded. On the other hand, in this case, for large z-scores the obtained figures of merit tend to the results obtained by the particle picking algorithm without screening process. The plots corresponding to the recall and *F*-measure are approximately planar for z-scores larger than 3. Observe that for a z-score threshold equals to 3, we obtain good results with a precision, recall and *F*-measure of 88.54%, 50.37% and 64.21%, respectively. Additionally, 93.30% of the particles are not disabled in this case.

In case of using a semi-supervised screening process (Fig. 7(b)), we have similar results for the precision than in the automatic case, but the recall is significantly increased and, therefore, also the *F*-measure. Observe that in this situation and for z-scores equal to 2 and 3, we obtain a precision, recall and *F*-measure of 90.68%, 51.55% and 65.70% and 88.35%, 51.70% and 65.23%, while the number of particles to be supervised is of 23.8% and 6.5% of the total (747), respectively. Note that with z-score thresholds of 2.0 and 3.0, and supervising the 23.8% and 14.19% of the particles, we obtain almost the same results than when we supervise every particle in the dataset. Additionally, observe that in both cases, the *F*-measure result is better that the one obtained when every particle of the dataset is manually supervised (fully-supervised).

Finally, we have also used the iterative processing option (automatic and semi-supervised) and we present the results in Fig. 7(d) and (e). Observe that, for the iterative case, we obtain similar results than in the automatic and supervised (Fig. 7(a) and (b)). The processing time required to perform the sorting and screening process in this dataset is of 56 s using a 2.4 GHz laptop in all of the cases presented before.

3.1.2. Second case

In the second experiment of this dataset, we previously picked 1219 particles using the same picking particle approach (Abrishami et al., 2013) than in the case before. As explained before, in this case, the particle picking algorithm obtains a small number of false negatives while a large number of false positives. The results obtained by the particle picking process without any particle screening are 70.27%, 84.75% and 76.83% for precision, recall and *F*-measure, respectively. Additionally, the results obtained when we supervise manually all the particles are 80.82%, 81.18% and 81%, respectively. In Fig. 8, we show the results of the proposed method for the automatic (a) and supervised (b) cases. Note that, as before, we have represented the precision and recall confident bands with red and blue colors, and the *F*-measure confident band is defined by the region between the two dashed black lines.

As can be seen from Fig. 8(a), in the automatic case, we obtain a precision rate close to the fully-supervised case (80.82%) when we use small z-score thresholds. However, the recall values are low because we also disabled a large number of true positive particles. Observe, that the figures of merit tend to the results obtained by the particle picking without screening process and the plots corresponding to the recall and *F*-measure are approximately planar for z-scores larger than 3, as in the case presented before. The results obtained for z-scores of 2 and 3 are 80.30%, 62.96%, 70.48% and 74.80%, 80.89% and 77.72% for the precision, recall and *F*-measure, respectively. The number of particles disabled for these z-scores are 33.60% and 7.90%.

In Fig. 8(b) we show the results obtained for the semi-supervised case. Observe that for z-scores of 2.0 and 3.0 we obtain 80.51%, 83.25%, 81.86% and 76.73%, 83.55% and 80.0% for precision, recall and *F*-measure respectively. In Fig. 8(c) we show the percentage of particles that has to be manually supervised. As can be seen from Fig. 8(c), the percentage of particles to be supervised for z-scores of 2.0 and 3.0 is of 24% and 9.8% respectively. Therefore, supervising only a small particle set, we obtain similar results than when we supervise every particle of the dataset (fully-supervised). Note that we do not show the results of the iterative processing because they are similar to the automatic and supervised cases. The required processing time for this dataset is of 48 s using the same machine than in the case before.

3.2. Human adenovirus type 2 (Ad2 ts1)

We have also checked our proposed method with another dataset acquired in our laboratory from a Human adenovirus type 2 (Ad2 ts1) samples, examined in a FEI Tecnai G2 FEG microscope operating at 200 kV and recorded on a Kodak SO-163 film under low dose conditions at a nominal magnification of 50,000× and digitized in a Zeiss Photoscan TD scanner using a step size of 7 μm (1.4 Å in the sample) (Pérez-Berná et al., 2009). One person has previously selected the particles manually, composing the reference or ground-truth set that is going to be used to obtain our results (figures of merit). The obtained particles have dimensions of 600 × 600 px and the dataset is composed of thirty micrographs. We have performed a fully manual screening process (fully-supervised) of this dataset by two different people. The results obtained by the first person after the fully-supervised process are 87.93%, 84.26% and 86.06% for the precision, recall and *F*-measure, respectively, and the total number of not disabled particles is of 1492. On the other hand, the results obtained by the second person are 88.19%, 83.94% and 86.01%, with a number of not disabled particles of 1483. From these values, we compute the fully-supervised results by the mean of these values that correspond to 88.06%, 84.10% and 86.03%, respectively. As before, we have used the Xmipp particle picking method (Abrishami et al., 2013) to determine particle locations. We have used conditions in which the output of the Xmipp particle picking method—input dataset of our post-picking sorting approach—is expected to have a relatively large number of false positives while a small number of false negatives (these conditions correspond simply to using a small training set of seventy particles from three micrographs). The results obtained from the particle picking task, without any screening process, correspond to precision, recall and *F*-measure of 77.38%, 85.48% and 81.23%, respectively, and the total number of detected particles is of 1720. In Fig. 9 we show the results obtained by our proposed method using the automatic (a) and semi-supervised (b) methods. Additionally, in Fig. 9(c) we show the percentage of particles that have to be manually screened by the semi-supervised approach.

As can be seen from Fig. 9(a), in the automatic case, we obtain a precision rate larger than the one achieved in the fully-supervised case when we use small z-scores, that are 88.47% and 88.07% for z-scores thresholds of 1.8 and 1.9 respectively. The recall values for these z-scores are 78.42% and 79.70% that still are very good results. Additionally, note that for a z-value of 2.4 we obtain a local maximum in the *F*-measure curve that corresponds to 84.97% that is very similar to the result of the fully-supervised case (86.03%). Finally, observe that the different curves are approximately planar for z-scores larger than 6 and converge to the results obtained from the particle picking task without any screening process.

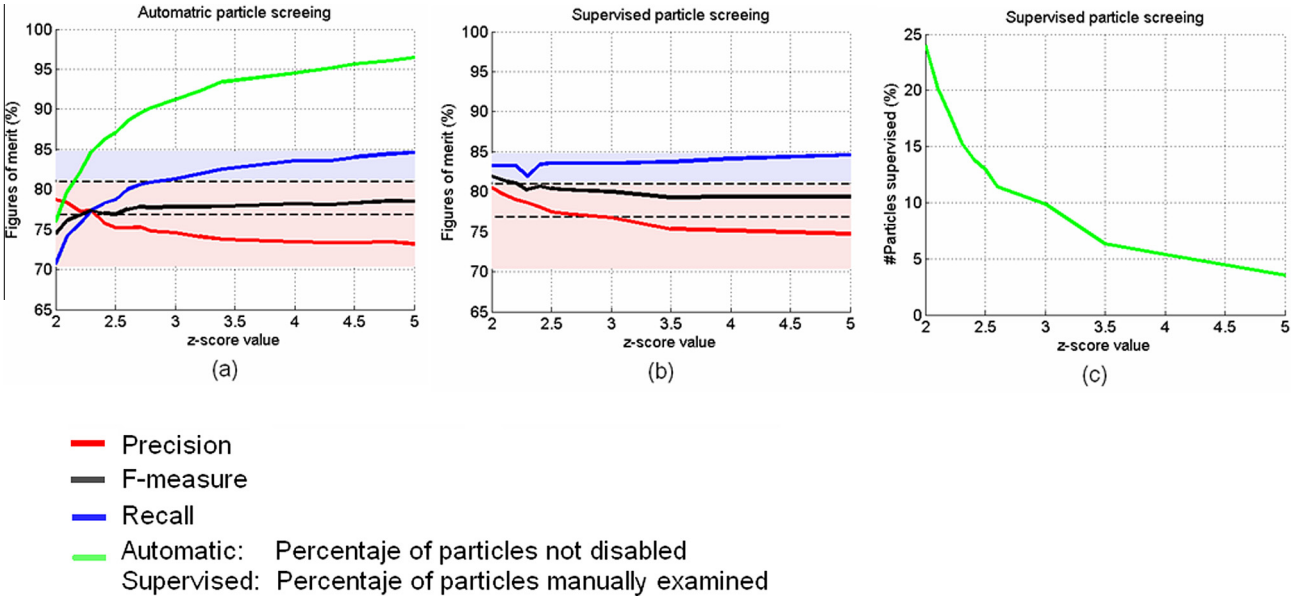

Fig.8. Figures of merit, where red, blue and black lines correspond to the precision, recall and *F*-measure rates, respectively, and obtained in the second case of the KLH dataset. Green lines are the percentage of particles taken into account (enabled) in the automatic case (a) and the percentage of particles supervised in the semi-supervised iterative case (b). Precision and recall confidence bands are represented with red and blue colors. The *F*-measure confidence band is defined by the region between the two dashed black lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

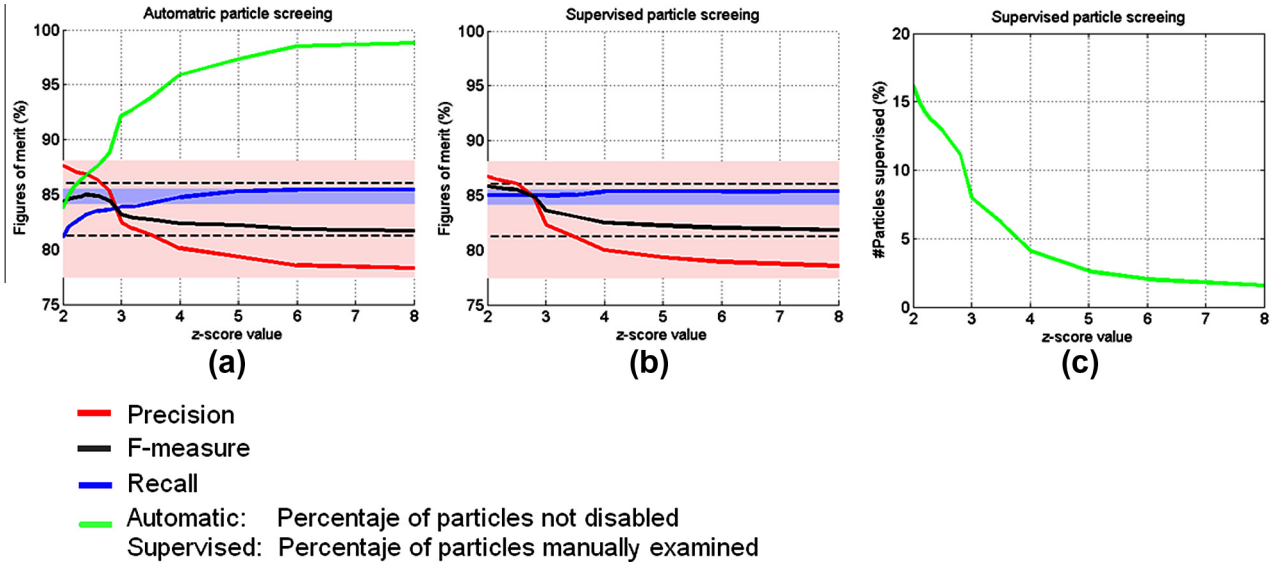

Fig.9. Figures of merit, where red, blue and black lines correspond to the precision, recall and *F*-measure rates, respectively and obtained in the second case of the Human adenovirus type 2 samples. Green lines are the percentage of particles taken into account (enabled) in the automatic case (a) and the percentage of particles supervised in the semi-supervised iterative case (b). Precision and recall confidence bands are represented with red and blue colors. The *F*-measure confidence band is defined by the region between the two dashed black lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The results of the semi-supervised case (Fig. 9(b)) show that the precision curve is very similar to the automatic one, but the recall is approximately planar with values very similar to the one obtained in the fully-supervised case. This causes the *F*-measure to be a monotonically decreasing curve. Observe that the *F*-measure values for z-scores smaller than 3 are very similar to the one obtained after the fully-supervised screening process. In Fig. 9 (c) we show the percentage of particles that are needed to be manually supervised. As can be seen from Fig. 9 (c), for low z-score values inside the range [1.8, 2.2] this percentage is defined between 19% and 14%, which corresponds to 232 and 171 particles to be supervised manually taking into account that the number of particles is of 1219. The number of particles disabled for these z-scores are 33.60% and 7.90%. The results obtained for a z-score threshold of 3 are 82.3, 85.0 and 83.6% for the precision, recall and *F*-measure respectively, which are very good results compared with the ones obtained by the particle picking task, without any screening process. In this case, the number of particles to be supervised is of 7.96%. Finally, the processing time required to compute this dataset composed by 1219 particles with size of 600×600 px corresponds to 9 min using a 2.4 GHz laptop.

4. Discussion and conclusions

In this work, we have presented a novel and fast post picking particle quality assessment and sorting method that can be used to recognize and discriminate erroneously picked particles from correctly picked ones. The input of our algorithm is a list of detected particles by any particle picking approach. The algorithm is based on three different hypotheses: (1) the majority of input particles have a low resolution "common shape" so that those input images that do not conform with this "common shape" can be labeled as incorrectly picked particles, (2) "good particles" have statistical properties not similar to the noise presented in the micrograph, (3) the histogram of gray levels is also similar for correctly picked particles. The proposed method is based on defining different sets ofparticle descriptors, which are morphology-based, histogram-based and noise-based. All of these descriptors are shift and rotational invariant. The proposed approach is very user friendly and intuitive, since for the calculation of the z-score of each particle, the user does not need to specify any parameter, being the process totally automatic. Furthermore, we have studied the statistical distribution of our z-score, reaching the conclusion that under a set of simple hypothesis a z-score of 3 (or around 3) may be used as a good automatic threshold, and then, the method can be used in a fully automatic manner from the initial input of particles. We have shown in the results section that the selection of a z-score threshold of 3 have provided good results in different cases. Attending to this z-score value we can reject incorrect detected particles using fully-automatic, semi-supervised or iterative approaches. We have tested our proposed approach using two datasets composed by micrographs of Keyhole Limpet Hemocyanin (KLH) particles and of Human adenovirus type 2 (Ad2 ts1) samples in different situations. In order to validate our results, we have used the 3DEM Benchmark site (http://i2pc.cnb.csic.es/3dembenchmark), that provides a robust computational infrastructure capable of supporting automatic benchmarking of the particle picking task, among others. Results obtained by our method have been compared with the ones computed after a fully-supervised process and without any screening procedure. These results show that after our proposed automatic curation process, the obtained precision rate is close to the one retrieved after a fully-supervised case when using low z-scores. However, the recall rate is low in this situation. Consequently, this curation process is especially well suited for cases where we have available a large set of particles and it is not problematic to lose a small percentage of correct ones. Note that this percentage is usually around 10%. In cases where we have obtained a small set of particles and we need to use all the available correct particles, it is better to use our proposed semi-supervised approach. Note that in Xmipp we have developed friendly graphical user interfaces to perform this supervision procedure. In this work, the presented results after performing the proposed semi-supervised approach show that the obtained precision for different z-score values is approximately similar to the one in the automatic case. However, the recall results are similar to the one obtained from the particle picking process, without any a posteriori screening process, which is the best one. Note that the percentage of particles that has to be supervised is around 10–20% for z-scores between 2 and 3. This limited supervising process makes

this task less awkward and time consuming and, at the same time, reduces the probability of performing mistakes in the manual screening process.

Acknowledgments

The authors would like to acknowledge economical support from the Spanish Ministry of Economy and Competitiveness through grants AIC-A-2011-0638, BFU2009-09331, BIO2010-16566, ACI2009-1022 and ACI2010-1088 as well as postdoctoral "Juan de la Cierva" grant with reference JCI-2011-10185. C.O.S. Sorzano is recipient of a Ramón y Cajal fellowship and J. M. de la Rosa-Trevín is supported by a CSIC grant (JAE-Predoc).

References

Abrishami, V., Zaldivar, A., de la Rosa-Trevín, J. M., Vargas, J., Otón, J., Marabini, R., Shkolnisky, Y., Carazo, J. M., Sorzano, C.O.S., A pattern matching approach to the automatic selection of particles from low-contrast electron micrographs, *Bioinformatics*, 2013 (in press)

Adiga, P.S., Malladi, R., Baxter, W., Glaeser, R.M., 2004. A binary segmentation approach for boxing ribosome particles in cryo em micrographs. Journal of Structural Biology 145, 142–151.

Arbeláez, P., Han, B.G., Typke, D., Lim, J., Glaeser, R.M., Malik, J., 2011. Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. Journal of Structural Biology 175, 319–328.

Basu, G., Ray, K., Panigrahi, P.K., 2010. Random matrix route to image denoising arXiv: 1004, 1356.

Chen, J.Z., Grigorieff, M., 2007. SIGNATURE: a single-particle selection system for molecular electron microscopy. Journal of Structural Biology 157, 168–173.

Filzmoser, P., 2005. Identification of multivariate outliers: a performance study. Austrian Journal of Statistics 34, 127–138.

Gonzalez, R.C., Woods R.E., 2007. Digital Image Processing, Prentice Hall, three ed, The ISBN number 9780131687288.

Hall, R.J., Patwardhan, A., 2004. A two step approach for semi-automated particle selection from low contrast cryo-electron micrographs. Journal of Structural Biology 145, 19–28.

Huang, Z., Penczek, P.A., 2004. Application of template matching technique to particle detection in electron micrographs. Journal of Structural Biology 145, 29–40.

Kovesi, P., 1997. Symmetry and asymmetry from local phase. In: Proceedings of the Tenth Australian Joint Conference on, Artificial Intelligence, 185–190.

Kovesi, P., 2002. Edges are not just steps, In: Proceedings of the Fifth Asian Conference on Computer Vision, 822–827.

Krzanowski, W.J., 2000. Principles of Multivariate Analysis, revised ed. Oxford University Press, Oxford.

Kumar, A., Smith, B., Borgelt, C., 2004. Dependence relationships between gene ontology terms based on TIGR gene product annotations. In: Proceedings of the third International Workshop on Computational Terminology, 31–38.

Langlois, R., Frank, J., 2011. A clarification of the terms used in comparing semi-automated particle selection algorithms in cryo-em. Journal of Structural Biology 175 (3), 348–352.

Larkin, K.G., Bone, D.J., Oldfield, M.A., 2001. Natural demodulation of two-dimensional fringe patterns. I. General background of the spiral phase quadrature transform. Journal of the Optical Society of America A 18, 1862–1870.

Mallick, S.P., Zhu, Y., Kriegman, D., 2004. Detecting particles in cryo-em micrographs using learned features. Journal of Structural Biology 145, 52–62.

Müller, R., 2004. Random matrices, free probability, and the replica method. Proceedings of European Signal Processing Conference.

Norousi, R., Wickles, S., Leidig, C., Becker, T., Schmid, V.J., Beckmann, R., Tresch, A., 2013. Automatic post-picking using MAPPOS improves particle image detection from cryo-EM micrographs. Journal of Structural Biology 182 (2), 59–66.

Ogura, T., 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. Journal of Structural Biology 145, 63–75.

Ogura, T., Sato, C., 2005. Auto-accumulation method using simulated annealing enables fully automatic particle pickup completely free from a matching template or learning data. Journal of Structural Biology 146, 344–358.

Pérez-Berná, A.J., Marabini, R., Scheres, S.H., Menendez-Conejero, R., Dmitriev, I.P., Curiel, D.T., Mantel, W.F., Flint, S.J., San Martin, C., 2009. Structure and uncoating of immature adenovirus. Journal of Molecular Biology 392, 547–557.

Plaisier, J.R., Koning, R.I., Koerten, H.K., van Heel, M., Abrahams, J.P., 2004. TYSON: robust searching, sorting, and selecting of single particles in electron micrographs. Journal of Structural Biology 145, 76–83.

Rath, B.K., Frank, J., 2004. Fast automatic particle picking from cryo-electron micrographs using a locally normalized cross-correlation function: a case study. Journal of Structural Biology 145, 84–90.

Roseman, A.M., 2004. Findem—a fast, efficient program for automatic selection of particles from electron micrographs. Journal of Structural Biology 145, 91–99.

Roweis, S., 1998. EM algorithms for PCA and SPCA. Advances in Neural Information Processing Systems 10, 626–632.

Short, J.M., 2004. SLEUTH—a fast computer program for automatically detecting particles in electron microscope images. Journal of Structural Biology 145, 100–110.

Singh, V., Marinescu, D.C., Baker, T.S., 2004. Image segmentation for automatic particle identification in electron micrographs based on hidden markov random field models and expectation maximization. Journal of Structural Biology 145, 123–141.

Sorzano, C.O.S., Marabini, R., Velázquez-Muriel, J., Bilbao-Castro, J.R., Scheres, S.H.W., Carazo, J.M., Pascual-Montano, A., 2004. XMIPP: a new generation of an open-source image processing package for electron microscopy. Journal of Structural Biology 148, 194–204.

Sorzano, C.O.S., Recarte, E., Alcorlo, M., Bilbao-Castro, J.R., San-Martín, C., Marabini, R., Carazo, J.M., 2009. Automatic particle selection from electron micrographs using machine learning techniques. Journal of Structural Biology 167, 252–260.

Ströbel, B., 1996. Processing of interferometric phase maps as complex-valued phasor images. Applied Optics 35, 2192–2198.

Vargas, J., Sorzano, C.O.S., Quiroga, J.A., Estrada, J.C., Carazo, J.M., 2013. Fringe pattern denoising by image dimensionality reduction. Optics and Lasers in Engineering 51 (7), 921–928.

Volkmann, N., 2004. An approach to automated particle picking from electron micrographs based on reduced representation templates. Journal of Structural Biology 145, 152–156.

Wolf, M., Garcea, R.L., Grigorieff, N., Harrison, S.C., 2010. Subunit interactions in bovine papillomavirus. Proceedings of the National Academy of Sciences of the United States of America 107, 6298–6303.

Wong, H.C., Chen, J., Mouche, F., Rouiller, I., Bern, M., 2004. Model-based particle picking for cryo-electron microscopy. Journal of Structural Biology 145, 157–167.

Zhang, X., Zhou, H.Z., 2011. Limiting factors in atomic resolution cryo electron microscopy: no simple tricks. Journal Structural Biology 175, 253–263.

Zhu, Y., Carragher, B., Mouche, F., Potter, C., 2003. Automatic particle detection through efficient hough transforms. IEEE Transactions on Medical Imaging 22 (9), 1053–1062.

Zhu, Y.B., Carragher, R.M., Glaeser, D., Fellmann, C., Bajaj, M., Bern, F., Mouche, F., de Haas, R.J., Hall, D.J., Kriegman, S.J., Ludtke, S.P., Mallick, P.A., Penczek, A.M., Sigworth, F.J., Volkmann, N., Potter, C.S., 2004. Automatic particle selection: results of a comparative study. Journal Structural Biology 145, 3–14.