International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:6, No:1, 2012

Automated Particle Picking based on Correlation Peak Shape Analysis and Iterative Classification

Hrabe Thomas, Beck Florian, and Nickell Stephan

***Abstract*—**Cryo-electron microscopy (CEM) in combination with single particle analysis (SPA) is a widely used technique for elucidating structural details of macromolecular assemblies at close-to-atomic resolutions. However, development of automated software for SPA processing is still vital since thousands to millions of individual particle images need to be processed. Here, we present our workflow for automated particle picking. Our approach integrates peak shape analysis to the classical correlation and an iterative approach to separate macromolecules and background by classification. This particle selection workflow furthermore provides a robust means for SPA with little user interaction. Processing simulated and experimental data assesses performance of the presented tools.

***Keywords*—**Cryo-electron Microscopy, Single Particle Analysis, Image Processing.

I. Introduction

IN recent years "single particle analysis" (SPA) matured to a key technology for structure determination in molecular structural biology [1]. The underlying principle is based on recording two-dimensional (2D), high-resolution electron micrographs each containing many, randomly oriented macromolecular complexes. By aligning and classifying the particle images iteratively the original projection angles can be determined and the three-dimensional density is formed by superposition of the "class averages". Ideally, the protein complexes are preserved during the data acquisition process in a close to native environment. This can be achieved by vitrification in a thin layer of ice, which comes at the cost that the sample is highly sensitive to radiation damage induced by the electron beam. Therefore electron micrographs are recorded under strict low-dose conditions and consequently suffer from low signal-to-noise ratios (SNRs) and low contrast levels [2]-[4].

Averaging of many particle images increases the SNR significantly and allows a meaningful classification and reconstruction of the underlying three-dimensional protein complex [1]. Typical particle numbers needed for this procedure reach from several tens of thousands to several millions. The extraction of the particle sub-images from the full electron micrographs is the initial processing step (Fig. 1) that is often still carried out by a manual, user- interactive selection procedure. With the advent of high- throughput data-acquisition routines [5][6], the need for automated particle extraction procedures became obvious and several strategies have been suggested since [2] [7]-[9]. Given their relatively robust performance at low SNRs, the majority of particle picking methods to date are based on cross-correlation using templates matching [9]. However, cross-correlation based methods still suffer from false positive and negative matches by which candidate particle images are incorrectly accepted or rejected. High contrast features caused by sample preparation or contamination are often the reason for these wrong hits. Another problem is the intrinsically low SNR of cryo-EM micrographs typically ranging from 0.01 to 0.1 [1]. Hence, the refinement of cross- correlation approaches to provide fast and deterministic selection under low contrast and SNR conditions given minimal a priori knowledge (aiding in automation and applicability to a wide selection of particles) is essential.

Here we present an enhanced cross-correlation based particle-picking algorithm unifying the calculation of a “Fast Local Correlation Function” (FLCF) with correlation peak analysis based on “Peak to Sidelobe Ratio” (PSR) and “Second Order Correlation” (SOC) [10][11]. By combining each of the afore-mentioned criteria by a weighted, stochastic optimizer, a final correlation score is determined [12]. The final candidate set of particle images is further refined by an iterative removal approach of false positives that is based on Principal component analysis (PCA) and K-means clustering. We demonstrate that this approach reduces significantly the requirements of a priori knowledge and exhaustive training by using only a minimal training set. We further demonstrate fidelity of the presented procedure by evaluating both synthetic and experimental data.

II. Algorithms

A. Optimizing cross-correlation

Correlation based localization of particles is widely used in particle picking [13]. In the following, we introduce the mathematical concepts used in this manuscript and strategies for peak analysis of two-dimensional cross correlation maps [14]-[16].

1) Fast local cross-correlation

Cross-correlation search approaches are based on the comparison of a searched, template S and an electron micrograph I. The similarity measure is based on the

All authors are with the Max Planck Institute of Biochemistry, Munich, Germany. (Corresponding author N. Stephan: phone: +49.89.8578.2651; e-mail: stephan@nickell.de).

normalized cross-correlation coefficient having an interval of [-1,1] (where -1 indicates a perfect inverted copy of *S* and *1* a perfect match of two identical images). Cross-correlation functions can be quickly calculated for large images by multiplying the Fourier transform of *I* with the complex conjugate of the Fourier transform of *S* [10]. The inverse transformation of the product yields cross-correlation coefficients at every position within the micrograph for *S.*

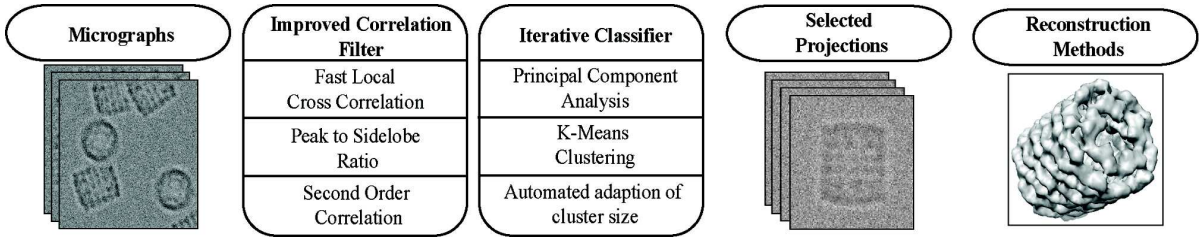

Fig. 1 Workflow of single particle analysis from microscopy to the resulting 3D protein structure
In our case, the improved correlation filter automatically localizes and extracts projections of macromolecules from micrographs. Next, the iterative classifier refines this selection by automatically adapting a cluster acceptance threshold. The remaining projections are then used to determine the macromolecules 3D structure

This can be expressed as:

$$CC = F^{-1}(F(I)\cdot F(S)^*) \quad (1)$$

where *F* denotes the respective Fourier transform, "*" denotes the complex conjugate and F^{-1} indicates the inverse transform. In this operation, the search template *S* (which is smaller in size than the search image) must be zero-padded to the size of the micrograph, prior to calculating the transform. Multiplication of the two transforms is an element-wise multiplication of the two matrices *F(I)* and *F(S).*

Each correlation score CCi is normalized by the number of considered points P in the template (i.e. the number of non zero pixels in the template) as well as the local averages, the standard deviations of S and the local area of I under the footprint of S. The localized normalization of cross-correlation coefficient is

$$FLCC_i = \frac{\frac{1}{P}CC_i - \bar{S}\cdot\bar{I}}{\sigma_S\cdot\sigma_I} \quad (2)$$

where $\bar{I}$ and σ_I are the mean and standard deviation of values in the image and the search template respectively [10]. $\bar{I}_i$ and σ_I are the local mean and standard deviation of *I* under the footprint of template *S* for each position *i* (corresponding to individual cross-correlation scores CC_i) within the image. The mean values $\bar{I}$ are calculated by convolution of *I* and a binary mask *(M)* corresponding to the zero-padded border of the search template *S.* In reciprocal space, the calculation of the local mean for each position in *I* is expressed as:

$$\sigma_I = \sqrt{\frac{1}{P}M\otimes I^2 - \bar{I}^2} \quad (3)$$

where ⊗ represents the convolution of *M* and *I* in reciprocal space. The standard deviation of the local area s_I under the footprint of the template is calculated by:

$$\sigma_I = \sqrt{\frac{1}{P}M\otimes I^2 - \bar{I}^2} \quad (4)$$

The scalar values $\bar{S}$ and σ_S are calculated in the same way as $\bar{I}$ and σ_I, corresponding to the (P/2 + *1*)th value of the calculated convolutions. We later refer to FLCC as the determined coefficients by the FLCF function.

2) Peak to sidelobe ratio

Correlation peak shape analysis has been introduced as an additional constraint to cross correlation based particle picking [16]. Valid particles should yield sharp peaks similar to the optimal shape of a delta function (a peak of infinite value

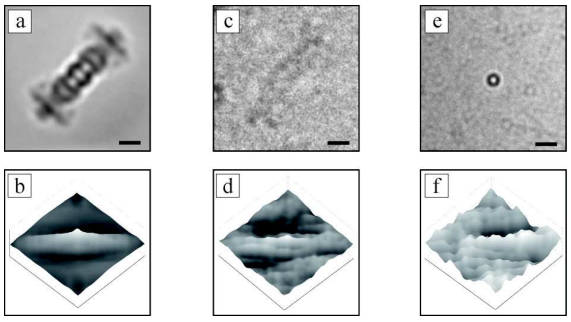

Fig. 2 The improved correlation filter analyses the peak shape to increase accuracy for identifying a template. (b) Autocorrelation-peak specific for template (a). (d) Reveals the correlation-peak-shape of (c) and (a). (d) resembles a damped copy of (b). (f) Depicts the peak of (e) correlated with (a). The pattern of (b) is not obtainable here. Thus, in order to improve picking reliability we extended correlation with two functions: one focusing on the peak sharpness (Peak to sidelobe ratio) and the other focusing on the peak shape (Second order correlation)

surrounded by zeros), similar to the autocorrelation of a template (Fig. 2). The peak to sidelobe ratio (PSR) is a measure of correlation peak sharpness, based on the central peak value relative to those of its neighbors and can be used to distinguish "sharp peaks" (indicative for valid particles) from "broad peaks" (indicative for false correlation maxima) [11].

The sidelobe *(SB)* is the region around a correlation maximum extending to the radius of the template image. This area can be increased by a factor of two for locating adjacent particles. A small, center-region around the peak is masked out since typical correlation peaks are wider than an ideal, single pixel [11]. Given the mean correlation coefficient $\overline{SB}$ of *SB* and the corresponding deviation $\sigma_{Sidelobe}$, the *PSR* for each peak in the correlation map is determined using the following expression:

$$PSR_i = \frac{FLCC_i - \overline{SB_i}}{\sigma_{Sidelobe,i}} \tag{5}$$

where $FLCC_{Map}$ are the all coefficients determined for every pixel in *I*. $\overline{SB}$ is calculated in a similar manner as $\overline{S}$

$$\overline{SB} = \frac{1}{P} SB \otimes FLCC_{Map} \tag{6}$$

The standard deviation $\sigma_{Sidelobe}$ is determined by:

$$\sigma_{Sidelobe} = \sqrt{\frac{1}{P} SB \otimes FLCC_{Map}^2 - \overline{SB}^2} \tag{7}$$

3) Second order correlation

Rather than only focusing on peak sharpness with the PSR, one may also focus on peak shape in order to further improve correlation fidelity to distinguishing the correlation profile of noisy image regions and unwanted image features from valid particles. By using Second Order Correlation (SOC), correlation peaks within $FLCC_{Map}$ are correlated with the autocorrelation function of the search template *(S)* (i.e. correlation of *(S)* with itself) using the FLCF.

$$SOC = FLCF(FLCF(I,S), FLCF(S,S)) \tag{8}$$

4) Training set

If no *a priori* information of the macromolecular structure is available, the user has to manually select a small number *(<100)* of particles for generating a training set *(TS).* This training set is aligned and averaged forming a template used for particle picking.

5) Optimized linear combination

The aim of combining the three methods (FLCC, PSR and SOC*)* was to improve correlation based particle selection. We merged the three functions to one by weighting *(a,b,c)* the respective function.

$$OLC = a \cdot FLCF + b \cdot PSR + c \cdot SOC \tag{9}$$

Thus, we optimized the linear combination so that highest values are assigned to true particles. For this purpose, representative points *p = (a,b,c)* are assessed according to an objective function *of(p)* using a training set *(TS).*

Given a list of candidate particles organized in descending order according to the corresponding OLC values, the objective function *of(p)* will yield a high score for *p* if the training set members have the highest OLC value. The next step was to find the global optimum in the feature space spanned by *a, b* and *c.* We determined the optimal linear combination for *(a,b,c)* using a standard stochastic optimizer, Simulated Annealing (SA)[17].

B. Iterative classification

Selection of particles with the above introduced OLC resulted in a reliable localization of candidate images *(CS)* on the micrographs. However, whilst localization guarantees a lower number of false negatives, the rate of false positives (showing up as outliers) turned out to be rather high. Hence, false positives need to be further classified and excluded from the candidate set.

The iterative classifier introduced here mimics the "human" particle picking process. Firstly, an experienced person would select the obviously correct particles, and would then increase the tolerance level by accepting particles having a slightly different appearance. Clear outliers are iteratively excluded by progressively reducing the maximal permitted entropy level between particles. False positives are thus removed gradually in an adaptive manner.

1) Principal component analysis and K-means clustering

Principal component analysis (PCA) is a tool used to analyze multi-dimensional data-sets in lower dimensions. Such a reduction of data-complexity limits the influence of noise by conserving data-typical features [8][18]. We used *PCA* on the union *U* of the training set *TS* and the candidate set *CS.*

$$U = TS \cup CS \tag{10}$$

This way, the number of true particles is increased and corresponding features are amplified significantly in the eigenimages. Other image features such as carbon, incoherent background and noise will be consequently mapped onto less significant eigenimages.

The union *U* is furthermore projected into a reduced space in a classical PCA manner [19], guaranteeing a more robust classification result. The number of eigenimages contributing to this reduction is regulated by the sum of their eigenvalues, which should at least cover 60% of the total variance. Experiments with values as high as 90% revealed the increasing influence of noise on classification.

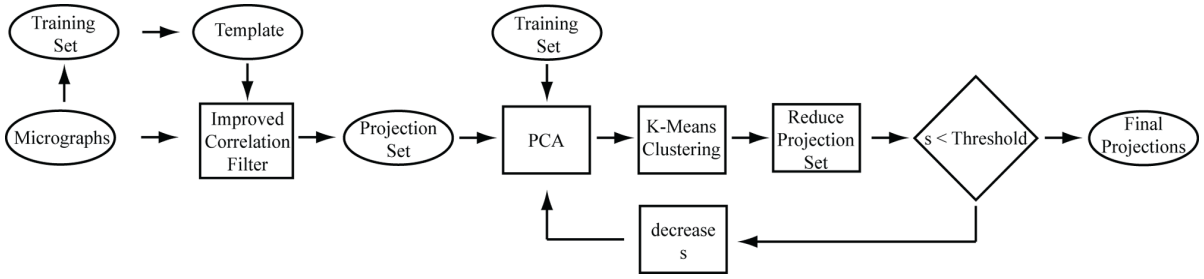

Fig. 3 Workflow of the particle selection algorithm. First, few projections in the micrographs are used to generate a Training Set from which a Template is compiled. Micrographs used here are excluded from later particle selection. The improved correlation filter robustly localizes projections similar to the template and yields a set of candidate projections. These projections are repeatedly decomposed with principal component analysis and clustered with k-means. Moreover, the constant training set supports features of the particle searched; projection set and cluster size is successively reduced to a user-defined deviation threshold of candidate projections. The remaining projections are eventually returned as the final projections

We then used K-means clustering as the second component of the iterative classifier to separate true particles from false positives [18]. Members t of TS are used to indicate k reference clusters for classification. We specify the cluster size s_j of cluster j by determining the mean member Euclidian distance μ_j from the center and the standard deviation σ_j.

$$s_j = \mu_j + \sigma_j \tag{11}$$

Every particle c in CS at a distance lower than d_{cj} from cluster centre j will be classified as a true particle (Fig. 3)

$$d_{cj} < s_j \tag{12}$$

2) Iterations

The workflow outlined above is processed iteratively (Fig. 3). Particles are extracted from the candidate set so that each new U yields eigenvectors different from those of the previous iteration. Thus, clustering is also repeated in each iteration.

We furthermore decrease s_{jm} for each iteration m with the use of a size factor sf:

$$s_{jm} = \mu_m + (sf_{Start} - m \cdot sf_{Step}) \cdot \sigma_{jm} \tag{13}$$

Hereby we reduce the accepted distance s_{jm} and, if required, decrease the probability to wrongly assign false positives to a cluster j. The values sf_{Start} and sf_{Step} are user dependent; typical values were $sf_{Start} = 3$ and $sf_{Step} = 0.25$. The iterative classification was stopped m_{End} when a cluster size of sf_{End} was reached:

$$sf_{End} = sf_{Start} - m_{End} \cdot sf_{Step} \tag{14}$$

III. Methods

A. Implementation

The algorithm described above was implemented in Matlab using the functionality of the TOM toolbox and the Matlab Distributed Computing Toolbox for parallel processing [20]. The parallel implementation of the particle picker uses a load balancing strategy that offers significant performance improvements, making a "real time" particle picking of individual micrographs feasible.

1) Simulated data

Test runs with simulated data provided a quantitative assessment of the developed algorithm whilst applying it to cryo-EM data demonstrated performance under "real world" conditions.

2) Signal to noise ratio of cryo-EM micrographs

For quantitative, representative testing of the algorithm under controlled, but realistic imaging conditions, simulated micrographs with a typical SNR of cryo-EM micrographs were generated. We applied the protocol from [1] to accurately measure the SNR in the micrographs. The values within the KLH dataset slightly varied around a value of $SNR=\sim 0.3$.

3) Simulated micrographs

Simulated micrographs were generated using randomly oriented particles of the (KLH) complex (model density taken from [7]). Each simulated micrograph contained ten side-views (used as true particles) and fifteen top-views (used as false particles) positioned randomly in the simulated micrograph. Overlapping side-views were removed from the list of true particles. Furthermore, an artificial carbon edge (typical for "real world" cryo-EM micrographs) was simulated. Image acquisition was modeled according to the following protocol [21]:

1. Orientation of side-views was randomized to mimic real electron micrographs.
2. An additive, Gaussian noise model was used for all simulated data. The standard deviation of the model was set to match the previously determined SNR estimates (Section 3.A.2). The SNR within each simulated micrograph varied (between 0.4 - 0.1) over different areas to simulate variations in ice thickness. (Fig. 4).
3. The contrast-transfer function (CTF) was applied to simulate the image acquisition process in a cryo electron microscope using a defocus setting of -3

μm and an acceleration voltage of 120 kV. A typical Modulation-transfer Function (MTF) of approximately 20% at 0.5 Nyquist was applied to the simulated electron micrograph.

4. Simulated micrographs were used to assess the localization by OLC. However, we also generated individual particles for testing the iterative classifier. Thus, we generated particle stacks of true particles (KLH side-views), false particles (KLH top-views) and particle stacks mimicking carbon and background noise. All particle stacks were generated using the same simulation procedure as described above (1-3).

B. Cryo-electron Microscopy Data

1) Keyhole Limpet Hemocyanin (KLH)

Although accurate simulated data allows quantitative testing and benchmarking of a newly developed algorithm, a standardized and widely accepted dataset comprising of many electron micrographs is required to demonstrate the actual performance under real-world imaging conditions (Fig. 4). Here we have applied our algorithm to the "KLH dataset", previously used in a particle picking bakeoff [7].

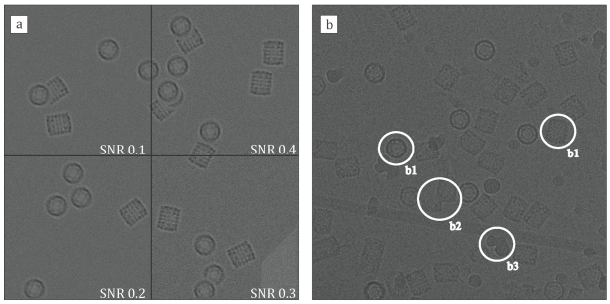

Fig. 4 Simulations of electron micrographs used for assessing the developed software. (a) Depicts artificial micrographs using side and top views of the Keyhole Limpet Hemocyanin (KLH) macromolecule with varying signal to noise ratios (SNR) within one image. Varying SNRs simulate varying ice thickness of vitrified samples. (b) Shows a micrograph taken from the KLH dataset used for benchmarking particle selection algorithms. Typical pitfalls for automated selection are low SNR, varying projection angles (b1), overlapping particles (b2) and (high-contrast) contamination (b3).

This dataset was collected at an electron microscope operated at an accelerating voltage of 120 kV using a 2048x2048 pixel CCD camera at 3μm under focus. The final magnification was 66,000x resulting in a pixel size of 2.2Å at the specimen level. Two reference lists of particle coordinates, one interactively picked by a user (Mouche) and one automatically generated (Selexon), were used as a control in this study [9][16].

2) 26S Proteasome

Additionally, a second dataset of the 26S proteasome was used for testing the algorithm [22]. In contrast to the KLH complex, micrographs of the 26S proteasome display a larger degree of structural heterogeneity due to the 26s lower stability. This results in potential false-positive hits. The ice-embedded 26S proteasomes were imaged using a Tecnai Polara electron microscope operated at 300 kV [22]. Micrographs were collected at an under-focus of 3.5μm with a GIF 2002 energy filter (Gatan Inc.). The final magnification of 82,500x yielded an object-pixel size of 3.6Å. An experienced user generated an interactively picked particle list of 18903 particles. The final three-dimensional reconstruction was computed using the XMIPP program package [23]. Resolution for these data set was determined to be 31.1Å using Fourier Shell Correlation at the 0.5 criterion [24].

IV. Results

A. Testing on Simulated Data

Both components of the picker, the optimized linear combination *(OLC)* (localization) and the iterative classifier (classification) were tested independently using the simulated data sets.

1) Localization

Table I lists the performance of the standard correlation function and OLC. Datasets comprised simulated micrographs with an invariant and a variant noise model (Section 3.A.2). We improved localization accuracy by using an extended peak shape analysis (OLC). OLC was determined to

$$OLC = 0.5 \cdot FLCF + 0.1 \cdot PSR + 0.4 \cdot SOC$$

Furthermore, ten true particles (Section 3.A.3) were present in each of ten simulated micrographs. Twenty candidate particles were selected from each simulation after picking.

2) Classification

The classification strategy was also tested for simulated data. However, here we used stacks of heterogeneous, aligned particles for testing. The complete particle stack consisted of true positives and false positives:

1. True positives – KLH side-views
2. False positives – KLH top-views / Background

TABLE I
False Positive *(FPR)* and False Negative *(FNR)* Rates Reveal a Higher Accuracy of the Optimized Linear Combination *(OLC)* compared to the Standard Correlation Function *(XCF)*

	XCF		OLC	
Noise	FPR	FNR	FPR	FNR
Invariant	60%	23%	62%	26%
Variant	62%	26%	56%	15%

Furthermore, we also generated a small training set to comply with the iterative workflow we presented (Section *2.B)* and merged it with the true positives. Hence, the final stack used for testing the iterative classifier consisted of ten training

set members, twenty true particles, thirty false positives and ten background images. Classification accuracies for different SNRs (Table II).

TABLE II
RESULTS OF THE ITERATIVE CLASSIFIER PROCESSED ON SIMULATED PARTICLE STACKS. PROCESSING WAS REPEATED FOR DIFFERENT SNRS TO DETERMINE THE PERFORMANCE UNDER VARYING CONDITIONS

SNR	FPR	FNR
0.5	0%	5%
0.1	2%	17%
0.01	13%	20%

B. Testing on Cryo-Electron Microscopy Data

1) Keyhole Limpet Hemocyanin

One hundred particles were interactively selected from a set of eight micrographs from the KLH dataset and were excluded from all further testing steps. A template *S* for picking with the improved correlation filter and a training set for the iterative classifier were generated from selected particles [25]. Best weighting coefficients of the three correlation methods were determined to:

$$OLC = 0.12 \cdot FLCF + 0.18 \cdot PSR + 0.70 \cdot SOC$$

Tuning of the iterative classifier was carried out manually using only few of the KLH micrographs as a pre-selected training set. An optimal configuration was found in a way that:

1. The first 5 eigenimages were used for projecting the particles into reduced space.
2. The training set was abstracted into 3 classes for generating the classification references.
3. The cluster size was reduced from $sf_{Start} = 3$ to $sf_{End} = 1.5$ in $sf_{step} = 0.2$

The performance of our algorithms was compared to a interactively (Mouche) and to an automatically picked reference (Selexon) (Table III).

TABLE III
RESULTS FOR THE BENCHMARK DATASET KLH. RESULTS OF THE DEVELOPED PARTICLE PICKER WERE COMPARED TO A INTERACTIVELY (MOUCHE) AND TO AN AUTOMATICALLY PICKED REFERENCE (SELEXON) BY ANALYZING THE RESPECTIVE FALSE POSITIVE AND FALSE NEGATIVE RATES

	FPR	FNR
Mouche	35.7%	18.7%
Selexon	32.4%	19.6%

2) 26S Proteasome

A second set of particle picking tests was performed using the 26S proteasome data. This was carried out by choosing an appropriate template and training set, generated from 500 manually selected particles. Optimal weighting of the three methods were determined by the optimizer and are given here:

$$OLC = 0.34 \cdot FLCF + 0.09 \cdot PSR + 0.57 \cdot SOC$$

For the 26S proteasome dataset, the iterative classifier settings were manually adjusted to the following configuration:

1. The first 8 eigenimages were used for projecting the particles into reduced space.
2. The training set was abstracted into 6 classes for generating the classification references.
3. The cluster size was reduced from $sf_{Start} = 3$ to $sf_{End} = 1.5$ in $sf_{step} = 0.25$

Here, we compared performance to an expert generated ground truth (Table IV). 51.3% of the automatically selected particles were in agreement with the expert list. Later inspection determined that 31.8% were incomplete components of the protein complex, contamination of imaging artifacts. Thus, 16.9% of the collected data were 26S Proteasomes that were not selected by the expert (Fig. 5).

TABLE IV
RESULTS OF THE AUTOMATED PARTICLE PICKER (15720 PARTICLES) COMPARED TO A GROUND TRUTH (18903 PARTICLES) GENERATED BY AN EXPERT

	Overlap	FPR	FNR
Expert	51.3%	31.8%	48.7%

V. DISCUSSION

An automatic particle selection algorithm was presented in this work, consisting of two main components: an improved correlation filter and an iterative classifier. For the first component, two novel approaches for peak shape analysis extend the classical correlation.

The optimal combination of the shape analysis methods and the cross correlation function was determined using a standard optimization algorithm. We found that this approach improved the fidelity of localizing protein complexes significantly.

An iterative classifier further refined the selection of detected complexes by iteratively repeating the PCA and the K-Means classification steps. Candidate particles were represented in a reduced data space where they align to predefined clusters, which were previously, determined using a small training sets of “true” particles. Conclusively, the combination of enhanced correlation with iterative classification and sorting of particles yields a robust and adaptive particle picking method. The complete workflow relies only on a small training set of 100 to 1000 particles for initialization that specifies expected positions of true particles in the feature space. Thus, one major benefit of this training scheme is that no large training sets of true / false particles as for Neural Networks or Support Vector Machines are required [26]. Both methods turned out to be a critical improvement towards a reliable and robust localization strategy for protein complexes in cryo-electron micrographs. Tests of the algorithms developed were carried out on simulated and real-world datasets.

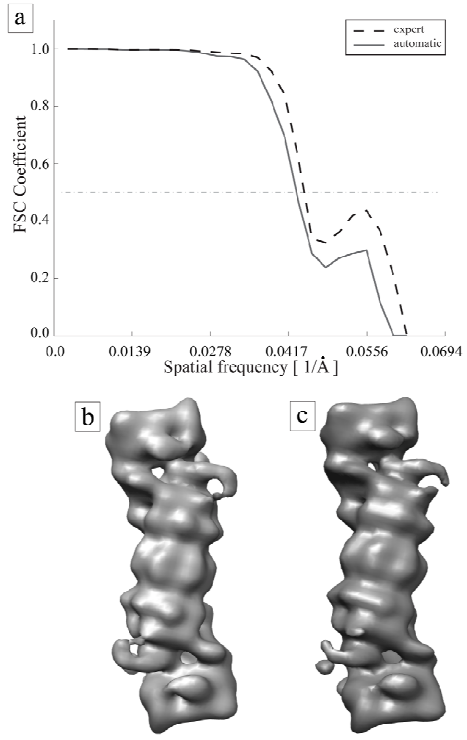

Fig. 5 26S Proteasome densities after particle selection. (a) Resolution as determined by the FSC=0.5 criterion for expert generated data (dotted line) and our automated particle picking algorithm (full line). (b) and (c) are the respective densities

The simulation protocol used included all imaging characteristics of cryo-electron micrographs such as CTF, MTF and noise. Both processing components were tested independently to determine their performance under varying imaging conditions.

Cryo electron microscopic data used for testing consisted of one established benchmarking dataset (KLH) and one data set of high interest for current research (26S Proteasome), respectively. Results on both revealed the method's reliability for particle picking, while making it comparable to other picking algorithms. Furthermore, the software processes the data in a parallelized manner, speeding up processing time for large datasets and even make the algorithm suitable for real time processing.

However, direct comparison of particle pick-lists to false positive / false negative are always problematic and hardly give accurate numbers in evaluating particle selection algorithms. In our experience, particle lists from trained experts already vary with error rates of up to 30% as we observed for the 26S proteasome dataset. The particle selection software developed here is furthermore part of the TOM toolbox [20] and is currently used for the analysis of the 26S Proteasome [27].

Acknowledgment

The authors thank Radosav Pantelic for carefully reading this manuscript. The research has received funding from the European Commission's 6th Framework Programme (grant agreement LSHG-CZ-2005-512028).

References

[1] Frank, J., Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. 2006: Oxford University Press.
[2] Adiga, P.S., et al., A binary segmentation approach for boxing ribosome particles in cryo EM micrographs. J Struct Biol, 2004. **145**(1-2): p. 142-51.
[3] Adiga, U., et al., Particle picking by segmentation: a comparative study with SPIDER-based manual particle picking. J Struct Biol, 2005. **152**(3): p. 211-20.
[4] Sander, B., M.M. Golas, and H. Stark, Advantages of CCD detectors for de novo three-dimensional structure determination in single-particle electron microscopy. J Struct Biol, 2005. **151**(1): p. 92-105.
[5] Nickell, S., et al., Automated cryoelectron microscopy of "single particles" applied to the 26S proteasome. FEBS Lett, 2007. **581**(15): p. 2751-6.
[6] Korinek, A., et al., Computer controlled cryo-electron microscopy - TOM(2) a software package for high-throughput applications. J Struct Biol, 2011. **175**(3): p. 394-405.
[7] Zhu, Y., et al., Automatic particle detection through efficient Hough transforms. IEEE Trans Med Imaging, 2003. **22**(9): p. 1053-62.
[8] Sigworth, F.J., Classical detection theory and the cryo-EM particle selection problem. J Struct Biol, 2004. **145**(1-2): p. 111-22.
[9] Zhu, Y., et al., Automatic particle selection: results of a comparative study. J Struct Biol, 2004. **145**(1-2): p. 3-14.
[10] Roseman, A.M., Particle finding in electron micrographs using a fast local correlation algorithm. Ultramicroscopy, 2003. **94**(3-4): p. 225-36.
[11] Kumar, B.V., A. Mahalanobis, and R.D. Juday, Correlation Pattern Recognition. 2005: Cambridge University Press.
[12] Russel, S.J. and P. Norwig, Artificial Intelligence. 1995: Prentice Hall.
[13] Nicholson, W.V. and R.M. Glaeser, Review: automatic particle detection in electron microscopy. J Struct Biol, 2001. **133**(2-3): p. 90-101.
[14] Caprari, R.S., Method of target detection in images by moment analysis of correlation peaks. Appl Opt, 1999. **38**(8): p. 1317-24.
[15] Volkmann, N., An approach to automated particle picking from electron micrographs based on reduced representation templates. J Struct Biol, 2004. **145**(1-2): p. 152-6.
[16] Woolford, D., et al., SwarmPS: rapid, semi-automated single particle selection software. J Struct Biol, 2007. **157**(1): p. 174-88.
[17] Duda, R.O., P.E. Hart, and D.G. Stork, Pattern Classification. 2. ed. 2000: Wiley-Interscience.
[18] Runkler, T.A., Information Mining. 2000: Vieweg.
[19] Forsyth, D.A. and J. Ponce, Computer Vision - A modern approach. 2003: Pearson Studium.
[20] Nickell, S., et al., TOM software toolbox: acquisition and analysis for electron tomography. J Struct Biol, 2005. **149**(3): p. 227-34.
[21] Forster, F., et al., Classification of cryo-electron sub-tomograms using constrained correlation. J Struct Biol, 2008. **161**(3): p. 276-86.
[22] Nickell, S., et al., Structural analysis of the 26S proteasome by cryoelectron tomography. Biochem Biophys Res Commun, 2007. **353**(1): p. 115-20.
[23] Scheres, S.H., et al., Image processing for electron microscopy single-particle analysis using XMIPP. Nat Protoc, 2008. **3**(6): p. 977-90.
[24] Saxton, W.O. and W. Baumeister, The correlation averaging of a regularly arranged bacterial cell envelope protein. J Microsc, 1982. **127**(Pt 2): p. 127-138.
[25] Short, J.M., SLEUTH--a fast computer program for automatically detecting particles in electron microscope images. J Struct Biol, 2004. **145**(1-2): p. 100-10.
[26] Bishop, C.M., Pattern recognition and machine learning. 2009: Springer.
[27] Bohn, S., et al., Structure of the 26S proteasome from Schizosaccharomyces pombe at subnanometer resolution. Proc Natl Acad Sci U S A, 2010. **107**(49): p. 20992-7.