An Improved Correlation Method Based on Rotation Invariant Feature for Automatic Particle Selection

Yu Chen[1,2], Fei Ren[1], Xiaohua Wan[1], Xuan Wang[3], and Fa Zhang[1]

[1] Key Lab. of Intelligent Information Processing
and Advanced Computing Research Lab., Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences Beijing, China
[3] Yanshan University, China
{chenyu,renfei,wanxiaohua,zhangfa}@ict.ac.cn,
wangxuan@ysu.edu.cn

Abstract. Particle selection from cryo-electron microscopy (cryo-EM) images is very important for high-resolution reconstruction of macromolecular structure. However, the accuracy of existing selection methods are normally restricted to noise and low contrast of cryo-EM images. In this paper, we presented an improved correlation method based on rotation invariant features for automatic, fast particle selection. We first selected a preliminary particle set applying rotation invariant features, then filtered the preliminary particle set using correlation to reduce the interference of high noise background and improve the precision of correlation method. We used Divide and Conquer technique and cascade strategy to improve the recognition ability of features and reduce processing time. Experimental results on the benchmark of cryo-EM images show that our method can improve the accuracy of particle selection significantly.

Keywords: Particle selection, Rotation invariant feature, Correlation, Divide and Conquer, Cascade strategy.

1 Introduction

Single particle cryo-electron microscopy (cryo-EM) has been widely applied to the study of macromolecular three dimensional (3D) reconstruction [1]. In cryo-EM, each biological sample must be prepared in a relatively homogeneous form, and then this sample is rapidly frozen as a thin film, transferred to the electron microscope and imaged, final, sufficiently sampled angular view of the 2D projections need to be aligned, combined to retrieve the samples 3D structure. However, to minimize radiation damage, micrographs showing projections of particles must be recorded at very low election dose, resulting in a high level of noise (the typical single-to-noise rate (SNR) is < 1) and very low contrast [2, 3]. To obtain atomic level reconstructed resolution, hundreds of thousands of particles

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 114–125, 2014.

may be necessary, which makes it impractical to manually pick the particles. In addition, particle detection by visual observation may be inaccurate and fairly subjective.

Several methods have been developed for automatic or semi-automatic particle detection [4]. Those algorithms,can be roughly grouped into two classes,template matching approach and feature-based approach.Feature-based approach usually relies on recognizing local or global salient features of particle images, such as statistical features [5,6], geometric features [7,8], cross-correlation features [9] or discriminative shape-related features [10].However, the main weakness of these methods is that it may be difficult to extract distinctive features from very low-contrast images. Template matching is a basic technique used in many signal processing and image analysis applications for detection and localization of patterns in signal corrupted by noise. In particle detection, this type of approach is based on cross-correlation using some templates (references) which are generated from either a 3D reference structure or the average of a few particles picked manually. It detects position of particle by comparing correlation coefficients of templates and target images.

Although higher accuracy results have been achieved by template-matching methods, comparing to feature-based approaches [11,12] , two key issues remain to be resolved. The first is how to improve precision by reducing the interference of noisy background when comparing correlation coefficients. As mentioned above, cryo-EM images usually have extremely low SNR, which will reduce the difference of correlation coefficients with templates between true particles and false particles(background or other type particles).The second is how to deal with the random orientation of particles more quickly. The orientation of particles is randomly in micrographs. To discriminate the orientation, templates must be rotated to generate a template set in different direction, and correlation with a template set is time-consuming especially when the number of target images is large.

To overcome these problems, we propose an improved correlation method based on rotation invariant feature to implement automatic and fast selection of particles. Rotation invariance of image means, for arbitrary rotation, function parameters may be changed, but the function value remained constant, and rotation invariant feature is a kind of feature that is shared by the same type particle in any orientation. Compared to conventional correlation-based methods, our method has several advantages.First,a preliminary particle set is selected using rotation invariant features, which contains about 98% target particles.Second, we apply correlation to filter the preliminary particle set instead of a whole cryo-EM image. Since the preliminary particle set contains less false particles compared to a whole image, filtering by correlation can further reduce the interference of noisy background and improve the precision of correlation method.Third,Divide and Conquer technique is applied in the extraction of rotation invariant feature to improve the recognition ability of features.Also a cascade strategy is used into the preliminary set generation to reduce processing time.

The remainder of the paper is organized as follows. In section 2, we introduce the framework of our method. First, we introduce the usage of rotation invariant features to generate a preliminary particle set and how the Divide and Conquer and cascade techniques are used to improve the processing. Then we introduce how to filter the preliminary set with correlation method. In section 3, we present our experimental results and analysis. Section 4 is focused on potential improvements.

2 Method

Our method consists of three steps(See Fig.1). The first step is to extract rotation invariant features to generate a strong classifier (discriminating true particles and false particles) by training. In this step, in order to improve the discrimination of classifiers, Divide and Conquer technique is used. The second step is to scan over the whole image with a particle size window and generate a preliminary particle set with classifiers. In this step, cascade strategy is used to speed up the recognition process. The third step is to filter the preliminary particle set with correlation function. In this step, mask is applied to every template and candidate particle before correlation to reduce the interference of background noise.

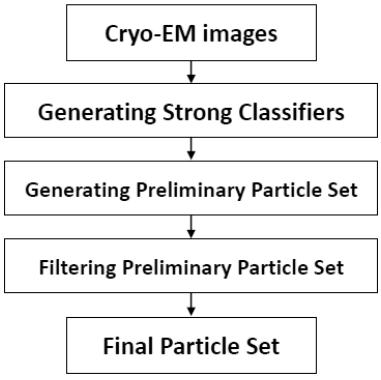


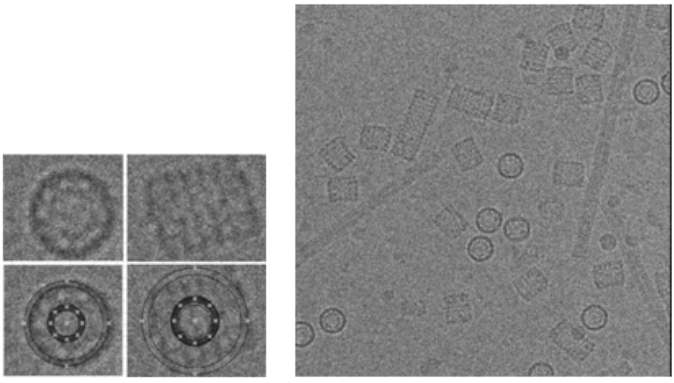
Fig. 1. Flow diagram of the improved correlation method

2.1 Rotation Invariant Features

Because of the random orientation of particles, it is difficult to generate some shared features in traditional way. However, if the center of a particle is determined, even if the orientation of the particle may be random, all the pixels of the particle are limited into a series of circles. And no matter how the orientation of the particle changes, every pixel just have a different position in the same circle. According to this, we generate some new features (see Fig.2(a)), which are invariant under rotation and well reflect the distribution of image gray, as

follows. Let $sum(r_0, r_1)$ is the sum of each pixel whose radius is more than r_0 but less than r_1, and SUM is the sum of all pixels of a particle size image. A feature is defined as Eq.(2.1).

$$f(r_0, r_1) = \frac{sum(r_0, r_1)}{SUM} \tag{2.1}$$



(a) Template and feature (b) raw cryo-EM image

Fig. 2. (a) templates and features of end and side-view (left and right,respectively); (b) raw cryo-EM image

2.2 Classifier Generation

After the first step discussed above, we generates a number of features. There is more than one reason to reduce the number of features to a sufficient minimum. Computational complexity is the obvious one. A related reason is that although two features may carry good classification information when treated separately, there is little gain if they are combined together in a feature vector. In this section, we describe a feature selection method which can select a set of the most important features from this huge feature space.

The Weak Classifier. For each feature f in feature space, we learn the training dataset and generate a classifier as Eq.(2.2).

$$C(f) = \begin{cases} 1 & , \quad \text{if} \quad (leftThr \leq f \leq rightThr) \\ 0 & , \quad \text{else} \end{cases} \tag{2.2}$$

Where $C(f)$ is the weak classifier for feature f. And $(leftThr, rightThr)$ is set to make $C(f)$ discriminate true particles and false particles correctly in most cases.

The Strong Classifier. We learn a set of training images and select only a few of optimal features for some weak classifiers. These selected features should have minimum mistakes to discriminate true particles and false particles in the training dataset. And we use a feature selection method to generate the strong classifier.

The feature selection method is described below:

1. Let $(x_1, y_1)...(x_m, y_m)$ where $x_i \in X = \{training \quad dataset\}$, $y_i \in Y = \{-1, +1\}, y_i = -1$ if x_i is false particle and $y_i = +1$ if x_i is true particle.
2. Initialize $D(i) = \frac{1}{m}, i = 1, ..., m$ as the weight of x_i.
3. Normalize as Eq.(2.3). And then for each feature $f_n, n = 1, .., N$, train a weak classifier $C(f_n)$. And the error of each classifier is given as Eq.(2.4).

$$D(i) = D(i) / \sum_{i=1}^{m} D(i) \tag{2.3}$$

$$e_n = \sum_{i=1}^{m} D(i)|C(f_n) - y_i| \tag{2.4}$$

4. Select T weak classifiers of lowest errors.
5. The strong classifier is the linear combination of these selected classifiers, as Eq.(2.5).

$$F(I) = \begin{cases} 1 & , \quad \sum_{t=1}^{T} a_t * C_t(I) \geq \tau * a_t \\ 0 & , \quad \text{else} \end{cases} \tag{2.5}$$

$$a_t = log(1 - e_t)/log(e_t) \tag{2.6}$$

Where a_t is the weight; τ is the threshold; I is a particle size image; top T features are used.

Divide and Conquer. The weight of each weak classifier is Eq.(2.6) in the strong classifier, but it might be not the optimal weight for each weak classifier. Thus Divide and Conquer technique is employed to optimize their weights and improve the recognition precision.

Firstly, we break the training dataset into several parts, and then, generate a strong classifier in each part respectively. Finally, we combine all strong classifiers into an integrated classifier with different weights as Eq.(2.7).The combination gives a larger weight to a weak classifier shared by two or more strong classifiers.

$$IntF(I) = \sum_{d=1}^{D} W_d * F_d(I), W_d = \frac{S_d}{S} \tag{2.7}$$

Where S_d is the image number of training dataset in dth part, W_d is the weight of the strong classifier F_d, S is the total image number of training dataset.

The Divide and Conquer technique is evaluated according to the performance of features which are trained with a training dataset containing 210 true particles and 210 false particles under different segmentation strategies. In the early stage,the recognition precision improves as the number of segments increases. However, too much division reduces training data in each divided part leading to an obvious deterioration.In this paper, we employ the features trained with a three-segment training dataset, of which each divided part contains at least 70 true particles and 70 false particles.

2.3 Preliminary Particle Set Generation

Threshold. As mentioned above, we need a threshold to control the performance of classifiers, which will influence the quality of the preliminary set. Generally, we evaluate a particle set with two attributes: the false positive rate(FPR)(See Eq.2.8) and the false negative rate(FNR)(See Eq.2.9).

If a proper threshold is used, one can get a preliminary particle set with acceptable FPR/FNR. Normally, we can obtain a proper threshold of one image by analysing the generated particle set, and then apply it to the whole image set captured under similar conditions. However, difference between proper thresholds of different images always exits, which is small but affects the accuracy, leading to a tradeoff between accuracy and labour.

If a low threshold is employed, one can get a preliminary set with about 98% target particles ($\sim 2\%$ FNR) but also with some false particles (high FPR), and then correlation can be used to filter the preliminary set to achieve an improved result.

$$FPR = \frac{the\ number\ of\ false\ selected\ particles}{total\ number\ of\ selected\ particles} \tag{2.8}$$

$$FNR = \frac{the\ number\ of\ missed\ true\ particles}{total\ number\ of\ true\ particles} \tag{2.9}$$

Cascade Strategy. As explained in the previous section, the strong classifier is a linear combination of several weak classifiers. Because most locations do not contain true particles, it is time-consuming to reject those locations by considering every weak classifier, and it is possible to reject those locations by only a few weak classifiers out of the selected ones [6]. Hence, we can employ a cascade of classifiers to speed up recognition. A cascade of classifiers groups all the selected weak classifiers in one strong classifier into several stages as shown in Fig.3. Each stage contains a few classifiers out of the selected weak classifiers. Most locations in the micrograph contain false particles are rejected at the early stage. The reason why the cascade of classifiers can speed up the detection process is that it removes the evaluation of other weak classifiers.

In this paper, we use three stages with 5, 10 and the other weak classifiers (about 20) respectively. The cryo-EM image is resized as $1024 * 1024$, and the processing time can be brought down to around 42 s from roughly 1254 s on a Intel(R) 2.33 GHz processor with threshold = 0.7.

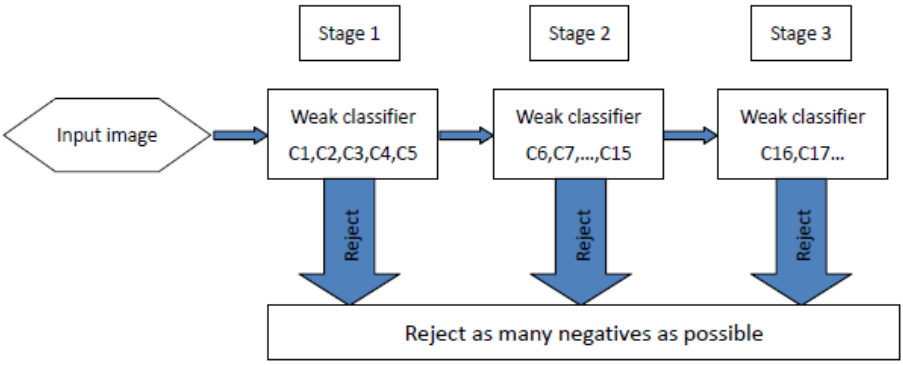


Fig. 3. A three stage cascade of classifiers

2.4 Filter the Preliminary Particle Set with Correlation

Template. Since dealing with a preliminary particle set instead of a whole cryo-EM image, we can simply use raw particle images as templates (see Fig.2(a)) and no image preprocessing is needed. However, because of the low SNR of cryo-EM images, some other views of particles, which should be considered as false particles, will also be selected into the preliminary particle set to become a disturbance especially when a low threshold is used. To deal with that, we should select some abtemplates those represent interfering views of particles.

Because of the random orientation of particles, all templates should be rotated, 5 degree per rotation, to generate a template set.

Correlation and Classify Candidate Particles. First, every template or candidate particle is masked with Eq.(2.10) to reduce the interference of background noise ,and then a fast Fourier based implementation of the correlation function is used to get correlation coefficients of candidate particles and templates/abtemplates.

$$I(x,y) = \begin{cases} I(x,y) & , \quad \sqrt{x^2+y^2} \leq R \\ 0 & , \quad \text{else} \end{cases} \tag{2.10}$$

Where I is a template or candidate particle, R is the radius of a particle.

According to the correlation coefficients of one candidate particle with all templates and abtemplates, we assign the candidate particle to the class represented by the template/abtemplate of highest correlation coefficient, which can provide orientation information for further research. And then we mark the score of this candidate particle as Eq.(2.11).

$$S(p_i) = \begin{cases} \max\{|coef(p_i,t_j)|\} & , \quad \text{assigned to template} \\ -1 * \max\{|coef(p_i,abt_j)|\} & , \quad \text{assigned to abtemplate} \end{cases} \tag{2.11}$$

Where $coef(p_i, t_j)$ is the correlation coefficient of the ith candidate particle and the jth template, and $coef(p_i, abt_j)$ is the correlation coefficient of the particle and the jth abtemplate.

After scoring all candidate particles, we can remove the candidate particles with low scores, which are usually false particles.

3 Results and Discussion

3.1 Test Dataset

Our method is tested with the keyhole limpet hemocyanin (KLH) dataset (available from the AMI group, The Scripps Research Institute, CA USA, http://ami.scripps.edu/ptrl_data/). KLH is a cylindrically shaped $\sim 8MD$ particle, a homo-oligomeric didecamer with D5 point group symmetry [13]. KLH particles are preferentially oriented into end-view and side-view (see Fig.2(a)). And only the selection of side-view is discussed in this paper.

82 digital micrographs of keyhole limpet hemocyanin (KLH) (see Fig.2(b)) are used. These images were acquired in a Phillips CM200 transmission electron microscope at a magnification of 66,000x and a voltage of 120kV. And they were recorded by a Tietz CCD camera of size $2048 * 2048$.

3.2 Preliminary Particle Set Generation with different Thresholds

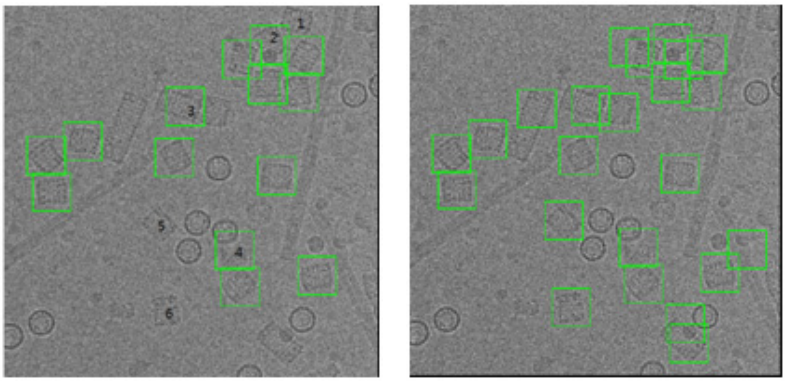
As mentioned in section 2, threshold influences the quality of the preliminary particle set. Next, we discuss the results of two representative cases.

Preliminary Set Generation with Proper Thresholds. Fig.4(a) shows the preliminary particle set of one image with a proper threshold. Although most true particles are selected with a proper threshold, there are some shortcomings of rotation invariant features should be improved. The selection of particle 2, 3 shows rotation invariant features are not sensitive to details of particles and easy to select some unqualified particles, increasing FPR; The selection of particle 4 and the miss of particle 5, 6 shows some noisy background area may satisfy rotation invariant features better than true particles.

For 82 images, with a proper threshold, rotation invariant features generate a preliminary particle set of 25.3% FPR and 22.2% FNR.

Preliminary Set Generation with Low Thresholds. Fig.4(b) shows the preliminary particle set of one image with a low threshold. All true particles are selected when some false particles are accepted wrongly.

For 82 images, with a low threshold, rotation invariant features generate a preliminary particle set of 69.4% FPR and 2.1% FNR; $\sim$ 98% of true particles are picked out.



(a) (b)

Fig. 4. Preliminary particle sets with different thresholds. (a) The preliminary particle set with a proper threshold. Threshold = 0.79, 14 candidate particles are selected containing 11 true particles. Particle 1 is rejected because it is too close to image edge; particle 2 has a small amount of contamination but still recognized as a true particle; particle 3 is part of a long particle but still accepted; particle 4 is noisy background but still satisfy high threshold; particle 5, 6 are true particles but missed. (b) The preliminary particle set with a low threshold. Threshold = 0.5, 23 candidate particles are selected containing 14 true particles.

3.3 Improved Correlation Method Based on Rotation Invariant Features

Fig.5(a) shows the final particle set of one image after filtering the preliminary set generated with a low threshold(see Fig.4(b)). There is only one false particle (particle 1) is left. Fig.5(b) shows the classification of 160 rectangular particles selected from 15 images.

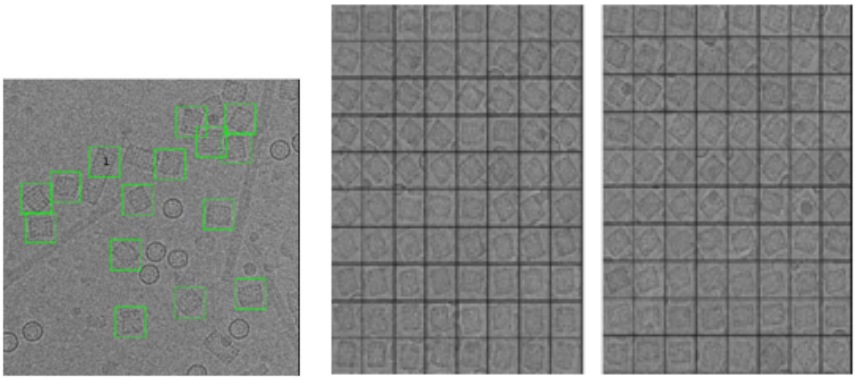
For 82 images, our method generates a final particle set of 12.89% FPR and 7.5% FNR.

Table.1 shows the comparison of our method and some other methods. The FPR of our method is 12.89% and higher than Sorzano's algorithm(9.3%), Volkman's algorithm(12.2%) and Malick's algorithm(11.7%), but compared to these three algorithms, our method has a lower FNR. The FNR of our method is 7.5% and only higher than Yu's algorithm(7.3%), but our method has a lower FPR. In summary, our method improves the accuracy significantly and generates one of the best results.

4 Potential Improvements

4.1 Accuracy Improvement

With correlation function, we filter a preliminary particle set to remove false particles and cut down the FPR (69.4% down to 12.89%). However, the FNR increases to 7.5%, which means we remove some true particles by error. Although



(a) (b)

Fig. 5. Results of improved correlation method. (a) : The final particle set of one image after filtering the preliminary set generated with a low threshold. Particle 1 is part of a long particle but left wrongly. Some edge-detection methods may be used to solve such problems (b) : The classification of 160 rectangular particles selected from 15 images.

$\sim 10\%$ FNR is acceptable and will not have a big impact on the accuracy of 3D reconstruction if we process more images to get enough particles, extra calculation is needed and slows down the method.

We can improve the FNR from two aspects. On one side, filtering methods can be applied to improve the SNR of cryo-EM images [14], as the quality of images improved, rotation invariant features will be improved to generate a better preliminary set, and better candidate particle images also means better performance of correlation function. On the other side, we can use traditional methods (like averaging templates to improve SNR) to generate better templates instead of raw particle images to reduce the increase of FNR.

However, both optimization methods mentioned above come with time consumption. It is important to balance speed and accuracy, especially in practical application.

4.2 Speed Up

As the requirement for precision increases, extra processing, like filtering, becomes necessary. Thus we should try to speed up the improved correlation method from other aspect.

After further evaluation, one can easily find out good parallel characteristics of our method. The method applies exactly the same process mode to every target image of particle size when generating the preliminary set, and for every candidate particle in the preliminary set, the same thing happens. Thus, it is possible to process each target image or candidate particle parallel on GPU, which will speed up the method in an amazing way.

Table 1. Comparison of rectangular particle selection between our method and other methods. It shows the FNR and FPR of detecting side views. The true dataset contains about 1042 particles picked manually by Fabrice Mouche.

Algorithm	FPR(%)	FNR(%)
Rotation invariant features with a proper threshold	25.3	22.2
Our improved correlation method	12.89	7.5
C.O.S. Sorzano et al.(2009)	9.3	30.9
Hall and Patwardhan (2004)	22.0	27.4
Ludtke(1999)	23.7	17.7
Volkman(2004)	12.2	27.4
Penczek(2004)	38.8	48.8
Yu and Bajaj (2004)	24.7	7.3
Malick et al.(2004)	11.7	14.2
Bern	16.2	23.8

5 Conclusion

In this paper, we propose an improved correlation method based on rotation invariant feature. Our focus is to solve two key issues of conventional correlation-based particle selection algorithms and improve the accuracy by introducing new rotation invariant features. The experimental results demonstrate that our method improves the accuracy of the particle selection significantly and the acceleration strategy used in the method is effective. We also discuss the potential improvements and good parallel characteristics of our method.

Our method is developed into software called Picker. Picker is freely available at our team's homepage http://ear.ict.ac.cn .

Acknowledgments. This work is supported by grants National Natural Science Foundation of China (61232001, 61202210, 61103139 and 60921002).The authors would like to thank Yangguang Shi,Fan Xu, Renmin Han,Jingrong Zhang for their helps about this work.

References

1. Joachim, F.: Three-dimensional electron microscopy of macromolecular assemblies. Academic Press (1996)
2. Henderson, R.: The potential and limitations of neutrons, electrons and x-rays for atomic resolution microscopy of unstained biological molecules. Quarterly Reviews of Biophysics 28(02), 171–193 (1995)
3. Sali, A., Glaeser, R., Earnest, T., Baumeister, W.: From words to literature in structural proteomics. Nature 422(6928), 216–225 (2003)
4. Zhu, Y., Carragher, B., Glaeser, R.M., Fellmann, D., Bajaj, C., Bern, M., Mouche, F., de Haas, F., Hall, R.J., Kriegman, D.J., et al.: Automatic particle selection: results of a comparative study. Journal of Structural Biology 145(1), 3–14 (2004)
5. Hall, R.J., Patwardhan, A.: A two step approach for semi-automated particle selection from low contrast cryo-electron micrographs. Journal of Structural Biology 145(1), 19–28 (2004)

6. Mallick, S.P., Zhu, Y., Kriegman, D.: Detecting particles in cryo-em micrographs using learned features. Journal of Structural Biology 145(1), 52–62 (2004)
7. Zhu, Y., Carragher, B., Mouche, F., Potter, C.S.: Automatic particle detection through efficient hough transforms. IEEE Transactions on Medical Imaging 22(9), 1053–1062 (2003)
8. Yu, Z., Bajaj, C.: Detecting circular and rectangular particles based on geometric feature detection in electron micrographs. Journal of Structural Biology 145(1), 168–180 (2004)
9. Sorzano, C., Recarte, E., Alcorlo, M., Bilbao-Castro, J., San-Martín, C., Marabini, R., Carazo, J.: Automatic particle selection from electron micrographs using machine learning techniques. Journal of Structural Biology 167(3), 252–260 (2009)
10. Abrishami, V., Zaldívar-Peraza, A., de la Rosa-Trevín, J., Vargas, J., Otón, J., Marabini, R., Shkolnisky, Y., Carazo, J., Sorzano, C.: A pattern matching approach to the automatic selection of particles from low-contrast electron micrographs. Bioinformatics 29(19), 2460–2468 (2013)
11. Roseman, A.: Findema fast, efficient program for automatic selection of particles from electron micrographs. Journal of Structural Biology 145(1), 91–99 (2004)
12. Sigworth, F.J.: Classical detection theory and the cryo-em particle selection problem. Journal of Structural Biology 145(1), 111–122 (2004)
13. Orlova, E.V., Dube, P., Harris, J.R., Beckman, E., Zemlin, F., Markl, J., van Heel, M.: Structure of keyhole limpet hemocyanin type 1 (klh1) at 15 å resolution by electron cryomicroscopy and angular reconstitution. Journal of Molecular Biology 271(3), 417–437 (1997)
14. Kumar, V., Heikkonen, J., Engelhardt, P., Kaski, K.: Robust filtering and particle picking in micrograph images towards 3d reconstruction of purified proteins with cryo-electron microscopy. Journal of Structural Biology 145(1), 41–51 (2004)