Journal of Structural Biology xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Journal of Structural Biology

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

Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection

Pablo Arbeláez [a], Bong-Gyoon Han [b], Dieter Typke [b], Joseph Lim [a,1], Robert M. Glaeser [b,*], Jitendra Malik [a]

[a] Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, USA
[b] Life Sciences Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 16 January 2011
Received in revised form 17 May 2011
Accepted 18 May 2011
Available online xxxx

Keywords:
Support vector machine
Texture analysis
Particle boxing

ABSTRACT

The goal of this study is to evaluate the performance of software for automated particle-boxing, and in particular the performance of a new tool (TextonSVM) that recognizes the characteristic texture of particles of interest. As part of a high-throughput protocol, we use human editing that is based solely on class-average images to create final data sets that are enriched in what the investigator considers to be true-positive particles. The Fourier shell correlation (FSC) function is then used to characterize the homogeneity of different single-particle data sets that are derived from the same micrographs by two or more alternative methods. We find that the homogeneity is generally quite similar for class-edited data sets obtained by the texture-based method and by SIGNATURE, a cross-correlation-based method. The precision–recall characteristics of the texture-based method are, on the other hand, significantly better than those of the cross-correlation based method; that is to say, the texture-based approach produces a smaller fraction of false positives in the initial set of candidate particles. The computational efficiency of the two approaches is generally within a factor of two of one another. In situations when it is helpful to use a larger number of templates (exemplars), however, TextonSVM scales in a much more efficient way than do boxing programs that are based on localized cross-correlation.

1. Introduction

In terms of the human effort required, manual selection ("boxing") of particles in electron micrographs is one of the most time-consuming steps in the process of single-particle structure analysis. Since images of 100,000 or more particles may be required for high-resolution reconstructions, it becomes increasingly important to use software tools to facilitate the boxing step (Glaeser, 2004; Nicholson and Glaeser, 2001). The questions that then arise include a concern whether the performance of one such method is better than another (Zhu et al., 2004), and whether the quality of an automatically boxed data set is as high as that of one generated by manual boxing. The precision (fraction of true positives in a data set) as a function of the recovery of true positives (termed "recall") could in principle be compared for different boxing programs, but in practice it is difficult to know whether a set of manually boxed particle images represents a comprehensive ground-truth labeling, i.e. whether such a data set itself has nearly 100% precision and recall.

In the current work, we present "TextonSVM", a new algorithm for automatically finding particles that is based on modern machine-learning techniques. This algorithm uses a support vector machine that is trained to recognize image-texture features. This algorithm thus provides an alternative to the widely used correlation-based techniques (Roseman, 2004), such as those used in SIGNATURE (Chen and Grigorieff, 2007). We further use unsupervised classification as a rapid and effective way to visually identify, and then eliminate, entire classes of unwanted false positives (Roseman, 2004; Shaikh et al., 2008) an operation that we refer to as "class-editing". We also use this class-editing approach, applied to the union of two or more independently boxed data sets, to estimate the comprehensive ground-truth labeling, which in turn enables us to use the "precision–recall" methodology as a second way to quantitatively compare different boxing methods. When using the Fourier shell correlation (FSC) function to characterize refined structures that are obtained from different data sets, we employ the more rigorous procedure of computing fully independent, 3-D density maps for arbitrarily chosen halves of a given data set (Grigorieff, 2000), rather than adopting the (still) common (but inappropriate) practice of dividing the data set before, or even after, the final iteration of alignment and Euler-angle assignments. As we explain in Section 4, we believe that the primary value of the FSC function is that it is sensitive, in a resolution-dependent way, to the relative heterogeneity of sets of particles that are boxed by different methods from the same set of digitized micrographs.

* Corresponding author. Fax: +1 510 486 6488.
E-mail address: rmglaeser@lbl.gov (R.M. Glaeser).
[1] Present address: Vision Group, Computer Science and Artificial Intelligence Laboratory, Massachussetts Institute of Technology, Cambridge, MA 02139, USA.

1047-8477/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2011.05.017

Results are presented for negatively stained specimens of four different macromolecular particles and for cryo-EM specimens of two of these particles. Negatively stained specimens were initially used to evaluate (1) the extent to which the performance of TextonSVM would meet or even exceed the performance of a well-regarded, correlation-based boxing method, and (2) the extent to which the relative performance of the two approaches might vary from one type of particle to another. Cryo-EM specimens (of two of these particles) were then used to evaluate the relative performance of the two approaches under conditions where improvements in automated boxing are more urgently needed but more difficult to achieve.

Three different data sets were generated for each of the negatively stained samples, one in which particles were boxed manually, one in which particles were boxed with TextonSVM, and one in which particles were boxed with SIGNATURE. In the case of cryo-EM samples, however, we used only data sets that were boxed automatically. When class-edited data sets were used, the FSC curves for particles boxed with different methods were almost indistinguishable for four of the six data sets, but for two data sets (negatively stained 70S ribosomes and unstained, cryo-EM RNAP II), the FSC curves for particles boxed with TextonSVM were systematically higher than the respective FSC curves for particles boxed with SIGNATURE. When the FSC function was used to compare reconstructions obtained from manually boxed particles to reconstructions obtained from automatically boxed particles, the quality of the automatically boxed data sets was again found to be essentially indistinguishable for three of the particles, but the improved quality of the negatively stained ribosome data set obtained with TextonSVM became even more apparent than before. In addition, we have evaluated the quality of data sets as they existed prior to class editing. As measured by our estimate of the precision–recall performance, we find that human performance is best, followed by that of TextonSVM, followed by SIGNATURE.

The results of our study demonstrate that the highly efficient combination of automated boxing and manual class-editing produces data sets that are as homogeneous as those obtained by manual boxing. This result increases our confidence that large data sets can be generated at a small fraction of the human effort that is required for manual editing of automatically boxed sets of candidate particles. In addition, we find that a texture-based boxing method that is based on modern computer-vision tools consistently performs at a level that is equal to or better than that of a cross-correlation based method.

2. Materials and methods

2.1. Sample preparation, electron microscopy, and data analysis

Four large protein complexes, representing a range of particle shapes and degrees of internal symmetry, were used as test specimens in this study. These particles, purified in the course of a comprehensive study of *Desulfovibrio vulgaris* Hildenborough (Han et al., 2009), are, respectively, lumazine synthase, an ~1 MDa icosahedral particle containing 60 copies of the 16.6 kDa riboflavin synthase β subunit; bacterioferritin, an ~480 kDa octahedral particle containing 24 copies of the 19.9 kDa protein subunit; RNA polymerase II, which was purified in this case as a dimeric, short-rod-shaped ~900 kDa complex thought to contain two copies each of core enzyme and NusA; and the 70S ribosome, an ~3 MDa, asymmetric particle with an irregular, globular shape. While all four specimens were used to make negatively stained samples, only two of them (lumazine synthase and the dimeric RNAP II–NusA complex) were used to make unstained, cryo-EM samples. Further information is given in Supplementary Material about the more standard aspects of electron microscopy and data analysis that were used, including manual boxing or particles; three-dimensional reconstruction and refinement; and use of the Fourier shell correlation (FSC) function to compare reconstructions.

2.2. Boxing particles with SIGNATURE

Manually boxed exemplars were masked with a circular mask that was just large enough to include an intact particle, in order to serve as the templates required as input to SIGNATURE. The rotational search was performed in 15° intervals over the full 360° range. The distance parameter was adjusted to be about 1.2 times the particle diameter in order to prevent overlapping particles from being picked. Using a few micrographs as a test, the peak-height threshold was adjusted to include about the same number of "true positive" particles per micrograph (after editing) as were picked by the texture-based program described below. The exemplars used as templates in SIGNATURE were the same as those, described below, that were used to train the texture-based program.

Automated boxing of particles in cryo-EM samples was performed with the same preprocessed, highly defocused (about 8 μm defocus) images, described below, that were used to box particles with the texture-based program. Phase flipping was applied for all the cryo-EM images, but no correction was applied to account for the amplitude of the CTF. The positions of the zeros in the CTF for each micrograph could be determined from the power spectrum of the image in our case, as samples were prepared on thin continuous carbon film.

The intermediate-defocus and low-defocus images were aligned against the highly defocused images by cross-correlation in order to transfer the particle coordinates to the former. These particle coordinates were then used to window out particles from the original micrographs by the program WI in SPIDER. Boxed particles were aligned by using the reference-free SPIDER command AP SR.

2.3. Boxing particles with a texture-based algorithm

2.3.1. Background

Texture is a primary perceptual cue used by humans when manually selecting particles in electron micrographs. Motivated by this observation, we developed a new algorithm for automatic particle detection called TextonSVM. Our approach models the appearance of images of molecular structures based on their texture, and, in addition, it applies a discriminative classifier called support vector machine (SVM) to differentiate particles from the background and from broken or otherwise unwanted particles. Since we have not described this approach previously, we do so here in some detail.

The concept of "textons" was introduced by Julesz (1981) to denote elementary units suitable for the analysis of textures in visual perception. In computer vision, an operational version of textons was proposed by Leung and Malik (2001), who convoluted an image with a bank of Gaussian filters to generate a vector of filter responses at each image location. These vectors were in turn clustered, and the vector at the center of each cluster was used as a texton. The entire set of textons then constituted a codebook that allowed one to represent the texture of new images. Textons can also be constructed by alternative ways, for example (as we do here) based on image patches rather than filter responses (Varma and Zisserman, 2003). Textons have been applied to multiple problems in computer vision and constitute what is currently the dominant approach for texture analysis.

Support vector machines solve a quadratic programming problem with linear constraints rather than by solving a non-convex, unconstrained minimization problem, as in standard neural

Please cite this article in press as: Arbeláez, P., et al. Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.05.017

network training. The method was originally introduced by Cortes and Vapnik (1995) and Vapnik (1995), where the problem was formulated as constructing the hyperplane that separates optimally two classes of data points. The notion of optimality focuses on the margin of (i.e. region close to) the decision boundary, and it is expressed in terms of the distances between data points and the boundary (hyperplane). The problem is thus written as a convex optimization, which allows to efficiently construct a globally optimal solution. The term SVM refers nowadays to a whole family of methods inspired by the original formulation, and these are generally considered to be some of the most successful machine-learning techniques for computer-vision applications. In this paper, we use Fast Intersection Kernel SVMs (Maji et al., 2008), which combine the computational efficiency of linear SVMs with the accuracy of non-linear classifiers.

Fundamental to our analysis of EM images, which generally have a very low signal-to-noise ratio, is the idea of describing a given image location by considering a local window centered at the pixel and taking into account the information within the neighborhood of that pixel, in addition to the pixel intensity itself. By varying the size of the window and the type of information considered, our framework allows to robustly address different tasks such as flat-field correction, denoising and automatic detection. In the following, we express the size of the square windows in terms of p, the diameter of the particle expressed in pixels.

2.3.2. Texture analysis

Our system learns the appearance of particles starting from a set of exemplars selected by the user. For this purpose, the exemplars are first extracted from the dataset and stored as individual images whose side is approximately $3p$. The number of exemplars was typically in the range 100–200, but in the case of cryo-EM images of RNA polymerase we used 1000 exemplars, since we found that this improved the performance of TextonSVM significantly. The number of exemplars used is thus likely to be something that one must experiment with for each new data set.

We first consider a patch of side about $p/4$ around each pixel of these images and represent it as a vector in $\mathbb{R}^m$ by concatenating the intensities of all the pixels in its interior. The vectors are then clustered using the k-means algorithm, with $k = 1000$ in the current experiments, and each cluster is represented by its center, which serves as a texton in our algorithm.

The set of textons forms a codebook that allows expressing each new image in terms of the texture of the exemplars. For this purpose, we consider a patch around each location in a new image and find its nearest texton in the codebook. We can then assign to the location both the label of this texton, $l \in [1, \ldots, k]$, and the value of its central pixel. After scanning all the locations in the new image, one obtains in the former case an image of labels, called a texton map, which we use for the detection stage. In the latter case, the result is a reconstruction of the original image.

Fig. 1 presents an overview of TextonSVM. In brief, we formulate particle picking as an exemplar-based learning problem, where we learn a model for the appearance of particles based on a given set of examples. When presented with a previously unseen image, our system uses the learned classifier to predict the probability of having a particle at each image location. The paragraphs below detail the different components of TextonSVM.

2.3.3. Pre-processing

As a consequence of the acquisition process, EM images often present unwanted intensity variations across the image and between different images. In order to standardize the appearance of particles in the dataset, our first stage of processing is to apply a local-normalization operation. For this purpose, given an input image I, we construct a normalized image N given by:

$$N(x,y) = \frac{I(x,y) - \mu_D}{\sigma_D}, \tag{1}$$

where (μ, σ) represent the mean and standard deviation, respectively, of I on a window D of side $3p$ centered at location (x, y).

After normalization, we extract the exemplars from the dataset and form the texton codebook mentioned above. We then produce filtered versions of the normalized images by assigning to each pixel the value of the central pixel in the corresponding texton. As can be observed from the especially challenging example shown in Fig. 3, the effects of this pre-processing operation are a flattening of the background, substantial reduction of background noise, enhanced visualization of particles, and greater uniformity of their appearance across the dataset. These properties are crucial for accurate automatic detection, even for more highly defocused images, for which visual detection of candidate particles is not as challenging as it is in this example.

2.3.4. Detection

We formulate the problem of particle picking as a visual pattern-recognition task where, given a set of labeled data, we would like to predict a label for new instances. In our case, the two classes are isolated particles and background, the labeled data are the exemplars provided by the user, and the new instances are found in previously unseen micrographs. For this purpose, we need to describe the visual pattern and then use this representation for its classification.

2.3.4.1. Representation. A rotation-invariant representation is important to account for there being unknown, in-plane rotations of the particles. However, note that the textons constructed above do not satisfy this property. Therefore, for the detection stage, we describe each texton by its intensity histogram. Then, using the χ^2 statistic as a measure of dissimilarity between histograms, we perform hierarchical clustering in order to obtain a lower number (about 100 in the experiments) of rotation-invariant textons, which we denote as r-textons.

The data representation is illustrated in Fig. 2. In order to describe the appearance of a particle, we consider a neighborhood "N" of side p centered in an individual exemplar image and form a first feature vector v_1 by concatenating the histogram of r-textons and the intensity histogram of the denoised image on N. For the appearance of the background, we consider a crown "C" of side $2p$ surrounding N, and form a second feature vector v_2 by measuring the same histograms in C. Our final representation for the first class of data, isolated particles, is the concatenation of these two feature vectors, noted v_1–v_2. Intuitively, the first part describes the appearance of the particle and the second one the appearance of the background surrounding it. This is the visual pattern we would like to identify.

In order to train a discriminative classifier, we need also exemplars for the second class which, in our case, rather than 'background', is 'anything that is not an isolated particle'. We construct them from the exemplars of the first class, as is shown in Fig. 2. Given the representation of such a positive exemplar, v_1–v_2, we construct three different negative exemplars: v_2–v_2, v_1–v_1 and v_2–v_1 which represent: background regions, agglomerations of particles and regions between particles respectively.

2.3.4.2. Classification. As a result of the previous stage, we have represented the exemplar particles as vectors in a high dimensional space $\mathbb{R}^m$ and, for each of them, we have constructed three different vectors representing non-particles. Since these vectors are (concatenations of) histograms, we compare them using the intersection similarity:

Please cite this article in press as: Arbeláez, P., et al. Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.05.017

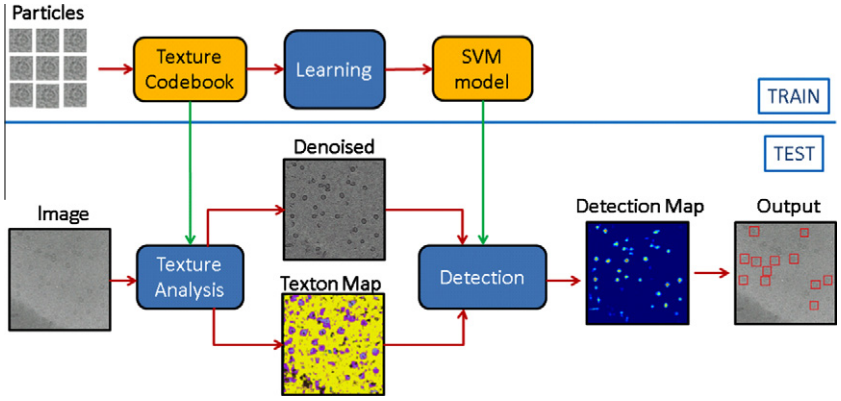

Fig.1. Overview of TextonSVM. At training time (top), exemplar particles are used to construct a texture codebook and to learn a model of particles with a non-linear SVM classifier. At testing (bottom), an image is first pre-processed by performing texture analysis, which produces a denoised image and a texton map. The classifier analyzes this appearance representation and estimates the probability of having a particle center at each image location (detection map). Finally, candidate particles are selected by extracting strong peaks of the detection map.

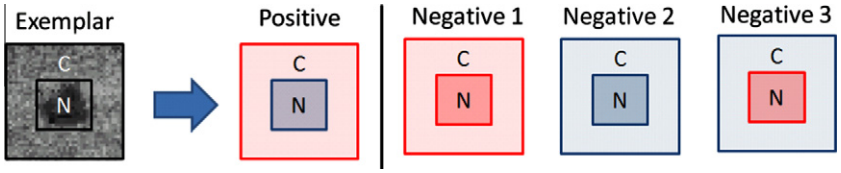

Fig.2. Data representation from particles provided by the user. For each exemplar, we consider a neighborhood covering the particle (N) and a crown surrounding the neighborhood (C). We represent each particle by considering its texture on the neighborhood (represented in blue) and the texture of the background in the surrounding crown (red). Our final representation for the positive class of data (isolated particles) is the concatenation of these two feature vectors. In order to train a discriminative classifier, we construct automatically examples for the negative class of data (non-particles) from the exemplars of the first class. These represent: background regions (e.g. "Negative 1" in which both N and C have a texture characteristic of the background); agglomerations of particles (e.g. "Negative 2" in which both N and C have a texture characteristic of the particles); and particles that are closer to one another than desired (e.g. "Negative 3" in which N has a texture characteristic of the background while C has a texture characteristic of the particles).

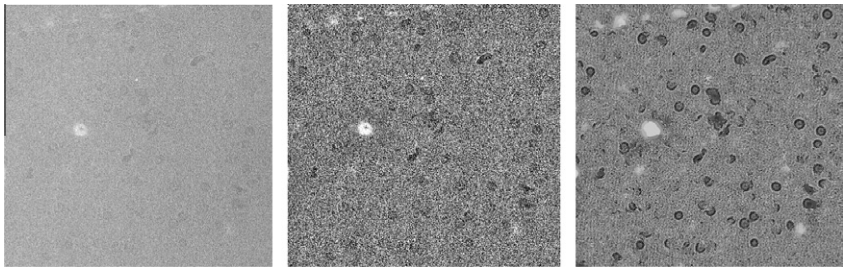

Fig.3. Improvement in particle visibility that results from local normalization of image data followed by texture-based denoising. The example shown here is for a cryo-EM image of lumazine synthase (MW ~1 MDa) that was recorded with an unexpectedly low defocus value of ~0.7 μm. This particular example is used to illustrate how effective the texture-based denoising algorithm is, even for cryo-EM images that are recorded at relatively low defocus values. (A) The digitized raw micrograph. (B) The corresponding image after local normalization. (C) The corresponding image after denoising.

$$K(u, v) = \sum_{i=1}^{m} \min(u(i), v(i)) \quad (2)$$

Then, we choose as classifier a support vector machine, which constructs the hyperplane in this space that best separates (in the sense of maximal margin) the two classes of data. The hyperplane splits the space in two halves, each of them representing one of the two classes.

Once the model is constructed, a new data point is classified by determining to which of the two half-spaces it belongs, and the distance from a data point to the decision boundary provides a measure of confidence in its belonging to one of the classes.

Given a previously unseen image, we evaluate the classifier in each location and assign to it the confidence provided by the SVM. The result is an image called a confidence map, which is an estimation of the probability of having a particle at each image location. As can be observed in Fig. 4, the peaks of the confidence map correspond to the particles. The final step of detection then consists in selecting the peaks that are above a user-defined threshold of confidence.

Optionally, the model can be improved by running the SVM on an image, labeling manually the false positives with high confidence, adding them to the set of negative exemplars and training a new SVM.

2.3.5. Optimization

Our system has a number of internal parameters that can be adjusted depending on the type of molecular structure considered, the most important ones being the exact size of the windows N and C defined above. When ground-truth is available, we propose to learn the values of parameters automatically with the following method:

Please cite this article in press as: Arbeláez, P., et al. Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.05.017

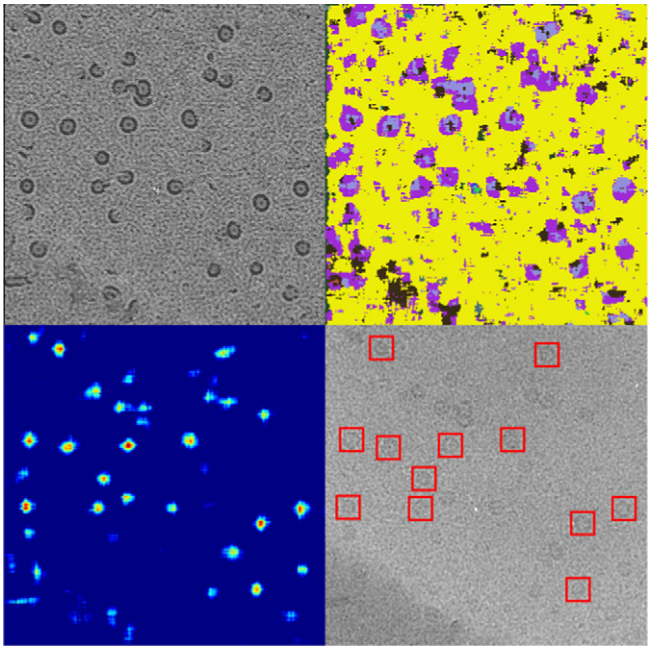

Fig.4. Key steps in the process by which TextonSVM boxes images of single particles. Top-left, denoised image; top-right, texton map; bottom left, confidence map; bottom-right, boxes overlaid on the original image, identifying particles detected at a confidence of 85%. The user can choose which confidence level to use for a given data set, so as to automatically box most of the particles that would be boxed manually, while at the same time boxing few candidate particles that would not be boxed manually, as is shown to be the case in this example. The example shown here is a cryo-EM image of lumazine synthase (MW ~1 MDa).

Consider one fully annotated micrograph, for instance, one from which the positive exemplars are extracted, and a fixed set of parameters. Then, we can run TextonSVM, evaluate the result with a precision/recall curve and use the area under the curve as a summary statistic of the performance with this set of parameters. Then, given a set of possible values for each parameter, we repeat the procedure for each combination and choose the ones with better performance.

Note that this grid search on the parameter space involves training and evaluating an SVM for each set of parameters and can be computationally expensive (in the current experiments, we consider about five choices for the sides of N and C). However, it optimizes the performance of the classifier and its cost is amortized over the number of images in the dataset, as it has to be done only once per type of particle. The resulting model can be evaluated on any number of new micrographs without adding computational burden.

2.3.6. User interface

We have developed a C++ stand-alone version of TextonSVM with a graphical user interface. The software runs on Linux platforms, and it is freely available to the research community at http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bio-images/textonsvm.

2.4. Class editing: the use of class averages to identify and remove "false positives" in automatically boxed data sets

Aligned particles were classified into about 300 classes by the program MSA in IMAGIC (van Heel et al., 1996). Class-average images that the user did not want to include in the final data set were identified by visual inspection. As an example, Fig. S1 shows a sorted data set of EM images of negatively stained RNAP II particles.

In the initial stage of removing unwanted particles one must proceed cautiously, especially if one has not yet established what the good particles should look like. It is thus useful to use at least two stages of classification and editing in which one first removes the most obvious "false positives" and then subjects the remaining data to a second round of unsupervised classification. This is because each cycle of editing can improve the subsequent classification due to the removal of outlier particles, thereby facilitating the identification of false positive particles in the next round.

2.5. Evaluation of the precision and recall of boxed data sets

When one or more fully annotated data sets are available, one can measure the performance of an automatic detection algorithm using precision/recall curves, a standard evaluation methodology in the information retrieval community (van Rijsbergen, 1979). Given a set of ground-truth particles and a set of automatic detections, their union can be divided in three parts: true positives (TP), false positives (FP) and false negatives (FN). Precision measures the amount of true positives among the detections and recall measures the fraction of ground-truth detected:

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad (3)$$

Precision and recall are two complementary measures of agreement of automatically detected particles with respect to ground-truth data. When the detector produces a single set of particles as output, one obtains a point in the precision/recall plane. When, as in the case of TextonSVM, the detector provides a measure of confidence, then choosing several thresholds produces a precision/recall curve, which describes its performance across all the detection regimes.

However, identifying the set of true particles in real experimental data poses a difficulty, since even manual (human) selection can be quite error-prone when the images are noisy. Thus, even a manually annotated data set may have both false positives and false negatives. In order to minimize errors of that type, we estimated the ground-truth set of particles by applying class editing to the union of sets of candidate particles obtained with TextonSVM, SIGNATURE, and – when available – manually boxed particles. This procedure has the advantage that it is blind as to whether a given candidate particle had been initially boxed by one or another of the methods. In addition, the improved signal-to-noise ratio that is present in class-average images reduces the ambiguity that one faces when deciding whether to include a particular image in the ground-truth data set.

3. Results

3.1. Automated boxing followed by class editing of negatively stained particles produces results similar to manual boxing, regardless of the type of particle

Visual comparison of the isosurface representations that are shown in Fig. S3 provides a convincing – although qualitative and subjective – demonstration that one obtains similar 3-D reconstructions of negatively stained protein complexes when using (1) manually boxed particles; (2) class-edited particles boxed with TextonSVM; and (3) class-edited particles boxed with SIGNATURE (Chen and Grigorieff, 2007). As was stated in Section 2, four different types of protein complexes were used in order to evaluate the performance of automated boxing software for particles with different shapes and different degrees of internal symmetry.

As is emphasized in Section 4, the one-dimensional Fourier shell correlation (FSC) curve provides a quantitative tool that is sensitive

Please cite this article in press as: Arbeláez, P., et al. Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.05.017

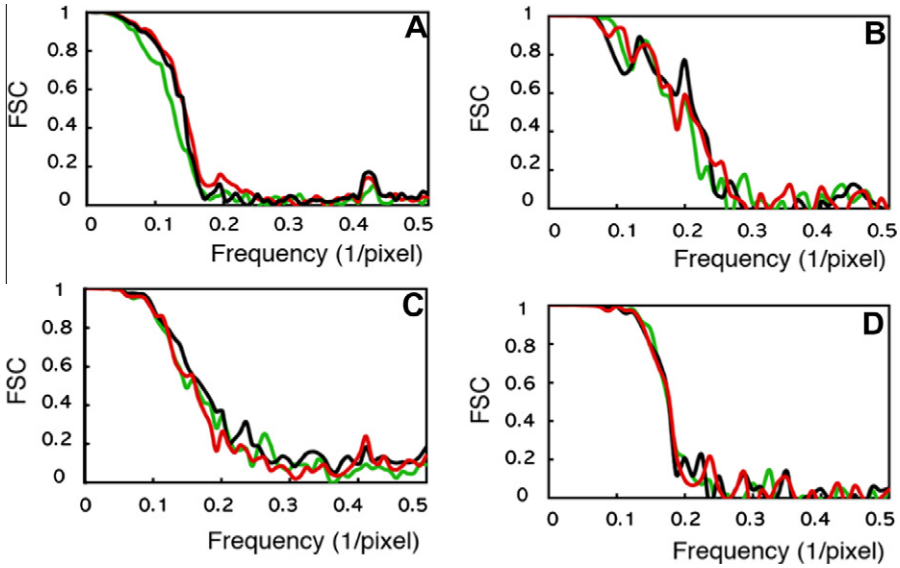

Fig.5. Use of the FSC curve to evaluate the resolution-dependent homogeneity of various data sets. Three different curves are shown in each panel in this figure, corresponding to manually-boxed data (black), data boxed automatically with SIGNATURE (green), and data boxed automatically with TextonSVM (red). In each case, the FSC function was used to compare the independently reconstructed and refined volumes that are shown in adjacent columns in Fig. S2. (A) 70S ribosome (MW ~3 MDa); (B) lumazine synthase (MW ~1 MDa); (C) dimer of RNA polymerase II complexed with NusA (MW ~900 kDa); (D) bacterioferritin (MW ~480 kDa).

to the resolution-dependent degree of homogeneity of different data sets that are derived from the same electron micrographs. We also emphasize that the shape of the entire FSC curve, and not just the resolution at which the curve falls to a given value (e.g. FSC = 0.5) provides information that is useful when comparing different data sets. Fig. 5 shows that manually boxed particles, class-edited particles boxed by TextonSVM, and class-edited particles boxed by SIGNATURE make up data sets that, in three of the four examples, have nearly identical properties, i.e. a similar degree of resolution-dependent homogeneity as judged by the similarity of the respective FSC curves. The one exception is the case of the 70S ribosome, for which there is evidence of greater heterogeneity in the set of class-edited particles boxed with SIGNATURE than is found in the set of manually boxed particles and in the set of class-edited particles boxed with TextonSVM.

A related, but different question is how homogeneous the data are when particles from one data set are compared to particles from another data set, for example manually boxed particles vs. class-edited particles boxed with TextonSVM or with SIGNATURE. FSC curves were therefore calculated for the refined structures obtained with class-edited particles boxed either with TextonSVM or SIGNATURE, versus the refined structures obtained with manually boxed particles. The resulting FSC curves are shown in Fig. 6 In the case of the 70S ribosome, it is now even more apparent than it was

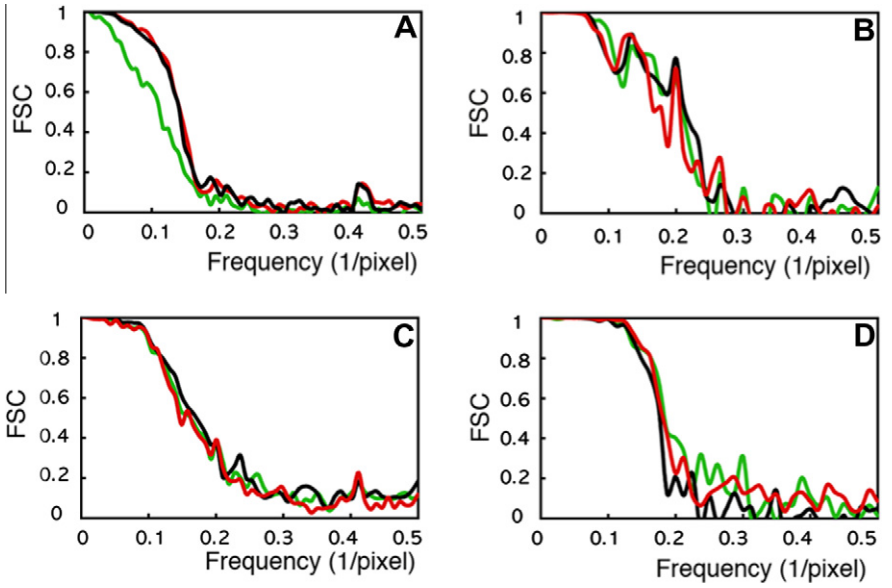

Fig.6. Use of FSC curves to compare the resolution-dependent similarity of volumes reconstructed and refined from data sets obtained by two different methods, one using automated boxing and the other manual boxing of particles. The alphabetical labeling of panels is the same as in Fig. 5. The green curves are for data sets obtained with SIGNATURE, while the red curves are for data sets obtained with TextonSVM. The black curves are the same "manual vs. manual" FSC curves that were shown previously in Fig. 5, which now are included in this figure as a point of reference with respect to the maximum homogeneity that can be expected within independent data sets for the respective specimens.

Please cite this article in press as: Arbeláez, P., et al. Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.05.017

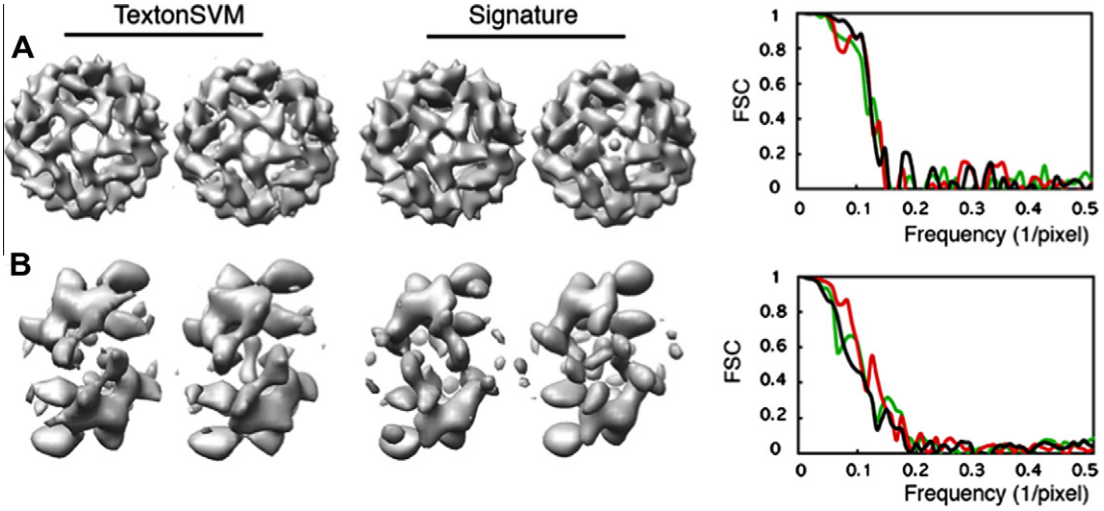

Fig.7. Comparison of results obtained with automatically boxed cryo-EM data sets. (A) Lumazine synthase (MW ~1 MDa). (B) Dimer of RNAP II complexed with NusA (MW ~900 kDa). The independently refined reconstructions that were obtained from separate halves of the full data sets are shown in the first four columns. The fifth column shows the FSC curves that were obtained by comparing the two volumes produced by particles boxed with TextonSVM, i.e. column 1 vs. column 2 (red); the two volumes produced by particles boxed with SIGNATURE, i.e. column 3 vs. column 4 (green); and one volume produced from the TextonSVM data set compared against one volume produced from the SIGNATURE data set (black).

in Fig. 5 that there is greater resolution-dependent heterogeneity between particles boxed manually and those boxed with SIGNATURE than is the case for particles boxed with TextonSVM. In the case of the other three molecular complexes, however, the particles boxed with either of the two automated packages were equally homogeneous relative to the manually boxed particles, as judged by the cross-data-set FSC curves. For reference, each panel also includes the FSC curves, which were already shown in Fig. 5, for manual vs. manual data subsets. These latter curves are included in each panel of Fig. 6 to serve as a standard that represents the best homogeneity that is currently achievable for micrographs of a particular type of particle.

3.2. Automated boxing of images of unstained particles together with class-editing produces similar FSC curves for TextonSVM and for SIGNATURE

A quantitative evaluation of the homogeneity of the edited cryo-EM data sets was performed, using the FSC function. Fig. 7 shows the refined reconstructions and the FSC curves for (A) lumazine synthase and for (B) the RNAP II–NusA complex, respectively. As before, data sets were divided into separate halves, and separate refinements were produced for each. FSC curves were then computed for the two volumes produced from particles boxed by TextonSVM (red curve) and for the two volumes produced from particles boxed with SIGNATURE (green curve). In addition, an FSC curve was computed for one volume produced from particles boxed by TextonSVM versus one volume produced from particles boxed by SIGNATURE. The three FSC curves are essentially indistinguishable for lumazine synthase, while the three FSC curves for the RNAP II–NusA complex show that the class-edited data set boxed with TextonSVM is slightly more homogeneous than the one boxed with SIGNATURE.

3.3. The precision–recall performance of TextonSVM is in all cases significantly better than that of SIGNATURE

The main difficulty in estimating the precision–recall characteristics of a given method of particle-boxing lies in obtaining an accurate approximation to the complete set of true positives. As implied above, manual boxing is prone to missing true positives through fatigue and inattention, and it is also prone to mistakenly accepting false positives due to the ambiguity of decisions that must be made when the noise level is high. In this regard, an advantage of employing class editing rather than performing particle-by-particle manual editing is that the noise level is reduced in class-average images, thereby reducing some of the ambiguity when decisions have to be made. We therefore regard data sets of class-edited particles that result from the union of different data sets to be a superior reference set that would allow at least a relative comparison of the precision–recall performance of different methods of boxing particles.

The full precision–recall curves obtained for images of different types of particles that were boxed with TextonSVM are shown in Fig. 8. Also shown in these graphs are single points representing the precision and recall for the data sets boxed with SIGNATURE, in which the user-defined parameters, including the threshold for accepting candidate particles, were the ones arrived at when generating the data presented above. In addition, the precision and recall achieved by manual boxing is also shown as another, single point on these graphs. Finally, a single point is also shown on the precision/recall curve for TextonSVM, which refers to the data sets that were first class-edited and then used to obtain the results shown in Figs. 5–7 and S2–S3.

The precision–recall curves produced for negatively stained particles by TextonSVM vary somewhat from one type of particle to another, but in general the curves are quite good. Except for the 70S ribosome, the precision ranges from 80% to well above 90% out to a recall of at least 80%, but the precision drops precipitously at a recall above 90–95%. The precision–recall curve for the 70S ribosome is generally somewhat below that of the other negatively stained particles, and it starts to drop steeply beyond a recall of about 75%. At the level of recall achieved during manual boxing, the precision of TextonSVM is very close to that of manual boxing for lumazine synthase and for bacterioferritin, but it is below that of manual boxing for our RNAP II complex and, of course, for the 70S ribosome.

The precision of boxing the same particles with SIGNATURE again varies with the type of particle, in the same way as it did for TextonSVM. At the same time the precision of SIGNATURE is consistently about 10% below that of TextonSVM, except in the case of the 70S ribosome particles, where the precision of SIGNATURE is about 25% below TextonSVM.

Please cite this article in press as: Arbeláez, P., et al. Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.05.017

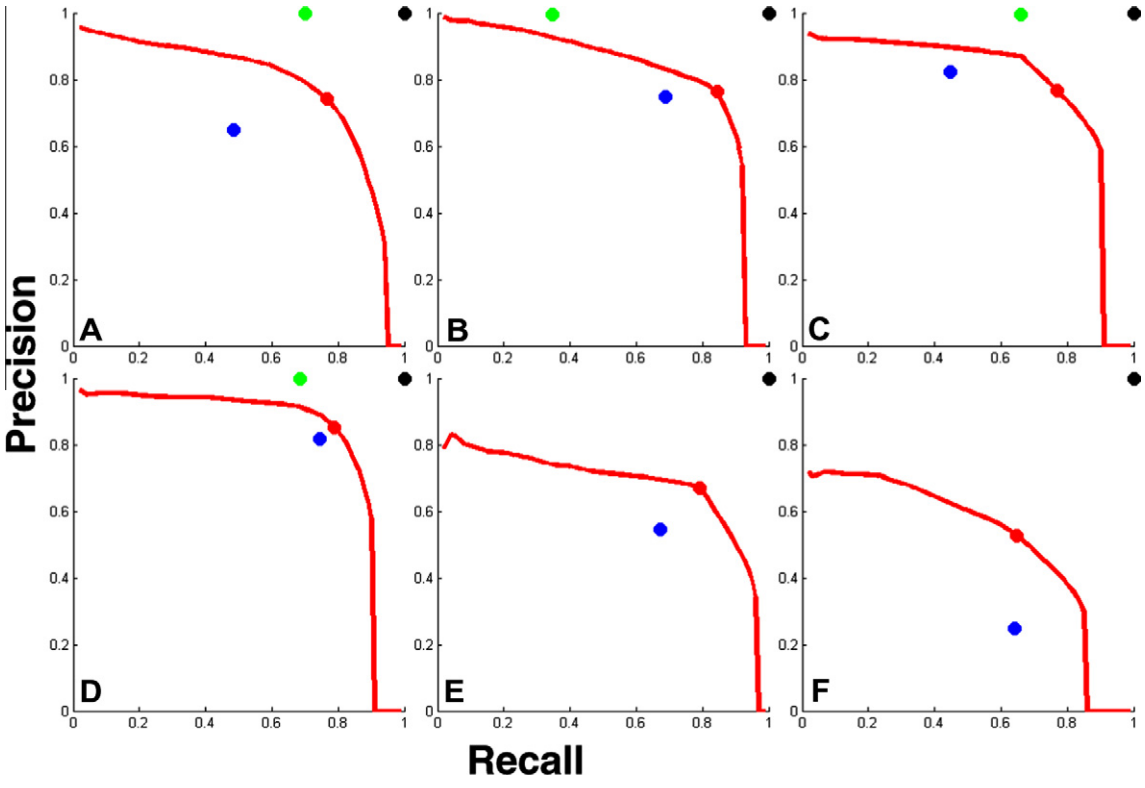

Fig.8. Evaluation of the precision–recall characteristics of TextonSVM, SIGNATURE (for the parameters we used in this study), and our own manual boxing (when applicable). (A) through (D) Images of negatively stained 70S ribosomes, lumazine synthase, dimers of RNA polymerase II complexed with NusA, and bacterioferritin, respectively; (E) cryo-EM images of lumazine synthase and (F) cryo-EM images of dimers of RNAP II complexed with NusA. In all cases, the reference data sets consisted of particles selected by class-editing the union of data sets obtained by TextonSVM, SIGNATURE, and (for negatively stained specimens) manual boxing. In all cases, the texture-recognition program, TextonSVM, provides data sets with better precision, at a given level of recall, than does the cross-correlation-based program, SIGNATURE. The red curve is the full precision–recall curve for particles boxed with TextonSVM, while the red dots correspond to the data produced as initial candidate particles, prior to class editing, by setting the confidence parameter to particular user-defined values for each data set. The blue dots correspond to the data produced by SIGNATURE, prior to class editing. The black dots correspond to the ground-truth data sets (with precision and recall both equal to 1.0 by definition) produced by applying class editing to the union of all available sets of candidate particles, while the green dots correspond to manually selected data sets.

In the case of cryo-EM samples, the precision of particle boxing by TextonSVM is significantly lower than it is for negatively stained particles, as might have been expected. At the same time, the precision achieved with TextonSVM is once again higher than that of SIGNATURE. Even in the case of TextonSVM, however, the precision achieved at almost any level of recall appears to be so low that editing is clearly desirable.

The raw micrographs used in this work, as well as boxed data sets and our best approximation to gold-standard data sets are available at http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bioimages/textonsvm.

3.4. Comparison of the computational efficiencies of TextonSVM and SIGNATURE

When TextonSVM and SIGNATURE are run on the same machine, and the original micrographs are processed at the same resolution, TextonSVM takes about two times longer per image to box particles than does SIGNATURE. In our tests, however, we have found that TextonSVM can be run with images that are subsampled by a factor of 2 relative to those used for SIGNATURE. In this case, TextonSVM runs about a factor of 2 faster than does SIGNATURE. Other considerations that can affect the computational efficiency include the fact that the computing time per particle scales linearly with the number of exemplars for SIGNATURE but remains constant for TextonSVM, since SIGNATURE matches each exemplar to a given image patch while TextonSVM makes a decision about where the patch is situated relative to the learned hyperplane. Our conclusion is that it is well-worth optimizing the parameters of either software package before boxing particles from large numbers of micrographs. When that is done, the run-time for boxing particles will not be a big consideration in choosing which software package to use for boxing.

4. Discussion

4.1. Our objective is to box, as closely as possible, the same particles that an investigator would box manually

Our perspective about automated boxing software is that it should reliably produce the same particles that investigators would box on their own. As a result, we generally do not expect automated boxing itself to avoid the bias that an investigator may or may not introduce when deciding what particles to include in the data set. On the contrary, our goal is to duplicate, as closely as possible, what the investigator would do, while greatly reducing the human effort required to complete that highly repetitive task. If, for example, the user chooses particles in a single orientation, intentionally or not, we expect the software to do the same. Similarly, if the user chooses structurally heterogeneous particles, we once again expect the software to do the same. Our perspective thus is that it is the responsibility of the user to initially choose exemplars that include particles in all desired orientations and conformations (or compositions), and then to subsequently edit the set of automatically boxed, candidate particles in order to remove any that the user would not have selected to begin with.

Please cite this article in press as: Arbeláez, P., et al. Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.05.017

4.2. Identification of candidate particles based on the texture of their images is an effective approach for automated boxing

The results obtained in our experimental tests with both negatively stained specimens and cryo-EM specimens demonstrate that texture-based identification of particles is an effective approach for automated boxing. In particular, we have found that the precision/recall characteristics of TextonSVM are consistently superior to those of SIGNATURE, which currently is one of the commonly used tools for automated particle boxing. In addition, our results indicated that the superior performance of TextonSVM may be especially helpful with difficult particles, such as the 70S ribosome.

The use of statistical measures of texture for automated particle boxing was investigated previously by Lata et al. (1995), who first identified candidate particles as peaks in a cross-correlation function that were higher than a user-defined threshold. They then assigned each such candidate particle to one of three categories: (true) particles, noise, and "junk", respectively. Our approach differs from that of Lata et al. considerably, both in our use of a high-dimensional measure, which – further – explicitly incorporates a pattern of "particle texture" surrounded by "noise texture", and in the use of a new generation of classifier (support vector machine) that is known to be superior to linear discriminant analysis both theoretically and in practice. In addition, we treat the problems of identifying candidate particles and classifying each candidate as "particle" or "non-particle" jointly, rather than as two sequential steps.

4.3. Use of unsupervised classification is an efficient way to edit automatically boxed data sets

A continuing shortcoming of all automated boxing programs is the fact that their precision/recall characteristics are not as good as that which is achieved by manual boxing. As a result, one must choose between setting the parameters for automated boxing to give a relatively high precision, at the expense of a rather low recall, or to set the parameters to values that give a relatively high recall, at the expense of a rather low precision. In the second case one must then edit the automatically boxed set of candidate particles in order to reduce the number of false positives to an acceptably low level. Manual editing is a poor choice, of course, in the sense that it largely defeats the purpose of automating the boxing process to begin with. In other words, the time required for editing remains a substantial fraction of the time to simply box particles manually.

Class editing, on the other hand, has a number of advantages in terms of the amount of human effort required. The number of images of particles that must be evaluated is reduced from the total number of particles in the data set (e.g. hundreds of thousands of particles) to the total number of classes that one asks the software to produce, typically something like 300. Even if one proceeds cautiously with the class-editing step, using two or more cycles of editing, the savings in time (human effort) is very substantial. We estimate, for example, that the human effort required to box 10,000 particles is already reduced from $\sim$20 h to $\sim$1 h by using an automated boxing program followed by class editing. Since much of the human time required for automated boxing and for class editing is independent of the size of the data set, the time-savings increases proportionally with the number of particles to be boxed, e.g. we estimate that a $\sim$200-fold reduction of human effort would result if the data set is to contain 100,000 particles. While these estimates are certain to depend upon the ability that one has to recognize true-positives in the images (or the class averages), which in turn will depend upon the experience of the user, we believe that these estimates are at least in the right order of magnitude.

While class editing has the further advantage that choices on which particle (classes) to keep and which to delete are made for less noisy images, it provides no help as regards the problem that user bias is introduced when making such choices. Indeed, the best that can be achieved is to produce the same data set that an experienced human would generate manually, but to do so with greatly reduced human effort.

The amount of computing time required for both the boxing step and the classification step is substantial, of course. The wall-clock time required for these computations depend greatly upon which particular computer hardware one has available, and one can be sure that these times will continue to decrease as the speed and parallelism of computational facilities increases. In general, of course, it is preferable to have a computer do most of the work, no matter how long it takes.

4.4. The use of two or more, independently boxed data sets may be especially useful when the particles are suspected of being "difficult to box"

The Fourier shell correlation (FSC) function (Harauz and van Heel, 1986; Saxton and Baumeister, 1982; van Heel et al., 1982) provides a resolution-dependent measure of how well two density maps, obtained from separate subsets of data, agree with each other. If the images within independent data sets represent projections of structurally identical particles, and if the data sets are large enough to provide well-sampled, uniformly distributed views of the structure, then identical procedures of reconstruction should produce identical density maps (at least as an idealized approximation). If, on the other hand, the data sets are heterogeneous in the sense that no single structure could exactly account for all of the images, then different 3-D reconstructions will be produced from independent data sets. In the first case, the FSC curve would approach 1.0 at all spatial frequencies (again, as an idealized approximation), but in the second case the FSC would be less than 1.0 by an amount depending upon how heterogeneous the data sets are within each resolution shell. When the number of particles included in the data sets is not limiting, one can thus infer that the shape of the FSC curve, and not just the point at which FSC = 0.5 (or any other point that is used as a criterion for "resolution") provides a resolution-dependent measure of the homogeneity (similarity) of the data sets. In this context we use the words homogeneity or heterogeneity, respectively, to describe the degree to which the 2-D images in a data set can or cannot be accounted for by a single 3-D object.

The results obtained in this study show that it can be worthwhile to compare the FSC curves for reconstructions produced by data sets generated by two independent methods of automatic boxing (e.g. TextonSVM and SIGNATURE). This may be particularly useful when one suspects that boxing a homogenous data set is especially difficult. In this case the use of the FSC curve for independently refined structures can show whether data sets obtained by quite different algorithms are equally limited by particle heterogeneity. In addition, as we have shown here, the FSC between structures refined from the independently boxed data sets can show whether the two methods are, indeed, boxing structurally similar particles. In the case of our sample of negatively stained 70S ribosomes, for example, we found that there was not only greater structural heterogeneity within the data set boxed by one software package (SIGNATURE) than within the data set boxed by a second software package (TextonSVM), but, in addition, there was greater heterogeneity between the two data sets than there was within either of them.

5. Summary and conclusions

We find that the use of a support vector machine to classify high-dimensional representations of image texture provides a

Please cite this article in press as: Arbeláez, P., et al. Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.05.017

superior method for identifying particles in both negatively stained specimens and unstained, cryo-EM specimens. In all six cases tested here, the precision–recall characteristics achieved with TextonSVM, a new software tool described in this paper, were found to be superior to those of a locally normalized, cross-correlation tool. Our tests included four examples of negatively stained particles whose molecular weights ranged from ~500 kDa to ~3 MDa, and two examples of unstained (cryo-EM) specimens whose molecular weights were ~800 kDa and ~1 MDa, respectively.

We also find that the improvement in precision, for a given recall, varies from one type of specimen to another (as well as with the chosen percent recall). In some cases the improvement was by as little as a factor of ~1.1, while in other cases it was by as great a factor as ~2. In spite of the improved performance of automated particle boxing that is achieved with TextonSVM, however, the precision often remains less than ~80% for a recall of ~70–80%. A higher precision at a recall of 80% still requires the use of a hybrid approach in which candidate particles are boxed with a lower threshold (i.e. confidence), followed by manual editing to remove unwanted particles. We confirm that the recently introduced method of class editing improves the precision of large data sets with minimal cost in terms of human effort.

Finally, we find that the quality of 3-D reconstructions that is achieved with class-edited data sets can still depend upon which method of automated boxing is used. It is still the case that TextonSVM is to be preferred over a correlation-based algorithm for particle picking because: (1) for our four negatively stained specimens we find that the FSC curves are essentially identical for data sets boxed manually or with TextonSVM, but when a cross-correlation algorithm is used, this is true for only 3 of the 4 specimens. (2) For our two cryo-EM specimens the FSC curves for automatically boxed particles is very similar, although one of them is again slightly better for the class-edited data set boxed with TextonSVM.

Acknowledgments

This work was performed as part of the research program of ENIGMA (http://www.enigma.lbl.gov), a Scientific Focus Area Program supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics: GTL Foundational Science through Contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the US Department of Energy. We are pleased to thank Dr. Ming Dong and Dr. Swapnil R. Chhabra for providing purified samples of the biological macromolecules used for this work.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jsb.2011.05.017.

References

Chen, J.Z., Grigorieff, N., 2007. SIGNATURE: a single-particle selection system for molecular electron microscopy. Journal of Structural Biology 157, 168–173.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.
Glaeser, R.M., 2004. Historical background: why is it important to improve automated particle selection methods? Journal of Structural Biology 145, 15–18.
Grigorieff, N., 2000. Resolution measurement in structures derived from single particles. Acta Crystallographica Section D – Biological Crystallography 56, 1270–1277.
Han, B.G., Dong, M., Liu, H.C., Camp, L., Geller, J., Singer, M., Hazen, T.C., Choi, M., Witkowska, H.E., Ball, D.A., Typke, D., Downing, K.H., Shatsky, M., Brenner, S.E., Chandonia, J.M., Biggin, M.D., Glaeser, R.M., 2009. Survey of large protein complexes in *D. vulgaris* reveals great structural diversity. Proceedings of the National Academy of Sciences of the United States of America 106, 16580–16585.
Harauz, G., van Heel, M., 1986. Exact filters for general geometry three dimensional reconstruction. Optik 73, 146–156.
Julesz, B., 1981. TEXTextons, the elements of texture, perception, and their interactions. Nature 290, 91–97.
Lata, K.R., Penczek, P., Frank, J., 1995. Automatic particle picking from electron micrographs. Ultramicroscopy 58, 381–391.
Leung, T., Malik, J., 2001. Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision 43, 29–44.
Maji, S., Berg, A.C., Malik, J., 2008. Classification using intersection kernel support vector machines is efficient. In: Computer Vision and Pattern Recognition, IEEE Conference, 2008, pp. 1–8.
Nicholson, W.V., Glaeser, R.M., 2001. Review: automatic particle detection in electron microscopy. Journal of Structural Biology 133, 90–101.
Roseman, A.M., 2004. FindEM – a fast, efficient program for automatic selection of particles from electron micrographs. Journal of Structural Biology 145, 91–99.
Saxton, W.O., Baumeister, W., 1982. The correlation averaging of a regularly arranged bacterial–cell envelope protein. Journal of Microscopy – Oxford 127, 127–138.
Shaikh, T.R., Trujillo, R., LeBarron, J.S., Baxter, W.T., Frank, J., 2008. Particle-verification for single-particle, reference-based reconstruction using multivariate data analysis and classification. Journal of Structural Biology 164, 41–48.
van Heel, M., Keegstra, W., Schutter, W., van Bruggen, E.F.J., 1982. In: Wood, E.J. (Ed.), Structure and Function of Invertebrate Respiratory Proteins. Harwood Academic, Reading, UK.
van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., Schatz, M., 1996. A new generation of the IMAGIC image processing system. Journal of Structural Biology 116, 17–24.
van Rijsbergen, C.J., 1979. Information Retrieval, second ed. Butterworths, London.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY.
Varma, M., Zisserman, A., 2003. Texture classification: are filter banks necessary? In: IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, pp. 691–698.
Zhu, Y.X., Carragher, B., Glaeser, R.M., Fellmann, D., Bajaj, C., Bern, M., Mouche, F., de Haas, F., Hall, R.J., Kriegman, D.J., Ludtke, S.J., Mallick, S.P., Penczek, P.A., Roseman, A.M., Sigworth, F.J., Volkmann, N., Potter, C.S., 2004. Automatic particle selection: results of a comparative study. Journal of Structural Biology 145, 3–14.