2005

Contributions To Automatic Particle Identification In Electron Micrographs: Algorithms, Implementation, And Applications

Vivek Singh
University of Central Florida

Contributions to automatic particle identification in electron micrographs: Algorithms, implementation, and applications

by

Vivek Singh
M.S. University of Central Florida, Orlando, USA, 2001
B.Tech. Indian Institute of Technology, Kharagpur, India, 1999

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the School of Computer Science
in the College of Electrical Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Summer Term
2005

Major Professors:
Prof. Dan. C. Marinescu and Prof. Hassan Foroosh

Abstract

Three dimensional reconstruction of large macromolecules like viruses at resolutions below 8 Å - 10 Å requires a large set of projection images and the particle identification step becomes a bottleneck. Several automatic and semi-automatic particle detection algorithms have been developed along the years. We present a general technique designed to automatically identify the projection images of particles. The method utilizes Markov random field modelling of the projected images and involves a preprocessing of electron micrographs followed by image segmentation and post processing for boxing of the particle projections. Due to the typically extensive computational requirements for extracting hundreds of thousands of particle projections, parallel processing becomes essential. We present parallel algorithms and load balancing schemes for our algorithms.

The lack of a standard benchmark for relative performance analysis of particle identification algorithms has prompted us to develop a benchmark suite. Further, we present a collection of metrics for the relative performance analysis of particle identification algorithms on the micrograph images in the suite, and discuss the design of the benchmark suite.

Dedicated to my parents.

On action alone be thy interest,
Never on its fruits.
Let not the fruits of action be thy motive,
Nor be thy attachment to inaction.

— *Bhagavad Gita, Chapter 2.*

Acknowledgments

I take this opportunity to thank individually, all those who have assisted me in one way or the other with my Ph.D work.

Dr. Dan C. Marinescu, my major advisor has been very helpful and supportive throughout the PhD program. But for his infectious drive to get things done, this work would be difficult to complete.

Dr. Hassan Foroosh, a co-advisor to me, for introducing me to nuances of many techniques of image processing, for helping me find directions whenever I had hit a roadblock and for patiently answering many of my *naïve* questions.

Dr. Sumanta Pattanaik for patiently attending my presentations and giving me many invaluable tips.

Dr. Huaxin You for introducing me to statistical data analysis techniques.

Dr. Fernando Gomez for pointing out the inadequacies of algorithms that were being developed and discussing improvements.

Dr. Yongchang Ji for patiently answer my questions regarding the macromolecular `3D` reconstruction programs.

Orit K. Schwartz for her help in developing the workbench.

The micrographs for testing of particle selection were obtained from **Dr. Tim Baker**'s lab at Purdue University. **Dr. Xiaodong Yang** had been extremely helpful in answering my incessant questions regarding biological aspects.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Progress lies not in enhancing what is, but in advancing toward what will be.

— *Khalil Gibran.*

Over the past two decades cryo-transmission electron microscopy (Cryo-EM) has emerged as an important tool, together with X-ray crystallography, NMR spectroscopy, and electron crystallography, to examine the three dimensional structures of macromolecules in their various conformational states. X-ray crystallography has been used for more than six decades to elucidate the three dimensional structures of biological macromolecules. The structure of DNA was discovered in part due to X-ray crystallography studies. The more recent techniques of NMR spectroscopy and electron crystallography also provide equivalent resolutions of 3D structures. However the following drawbacks inherent in these techniques limit their application.

1. Sample preparation times for these techniques, which run in days and months, are far longer than the few minutes required for preparation of samples for Cryo-EM.

2. Crystallographic techniques require preparation of crystals of the biological sample under investigation. However many of the interesting macromolecules are difficult to crystallize.

3. All three techniques do not work for large macromolecules (> 100 MDa).

4. Crystallization of samples does not allow for collection of structure related data of the samples in it's various conformational states.

The medium to low resolution (8Å to 30Å) structures of biological molecules and macromolecular complexes obtained from single particle technique can be complemented by the high resolution structures of their constituents obtained from X-ray crystallography, NMR spectroscopy, and electron crystallography. Cryo-EM specimen preparation procedures permit viruses and other macromolecules to be studied under a variety of conditions, which enables the functional properties of these molecules to be examined [2], [78], [82], sometimes dynamically [5]. This has resulted in an array of techniques for three dimensional high resolution structural study of macromolecules which was not possible before. Figure 1.1 displays a high level schematic of the complete procedure of obtaining a three dimensional density map of a macromolecule using Cryo-TEM.

1.1 Motivation and Goals

Structural studies for elucidation of three dimensional structure of macromolecules in their various conformational states allow biologists to better understand the workings of biological macromolecules and complexes. Currently there exists a a difference in resolution of about an order of magnitude between the struc-

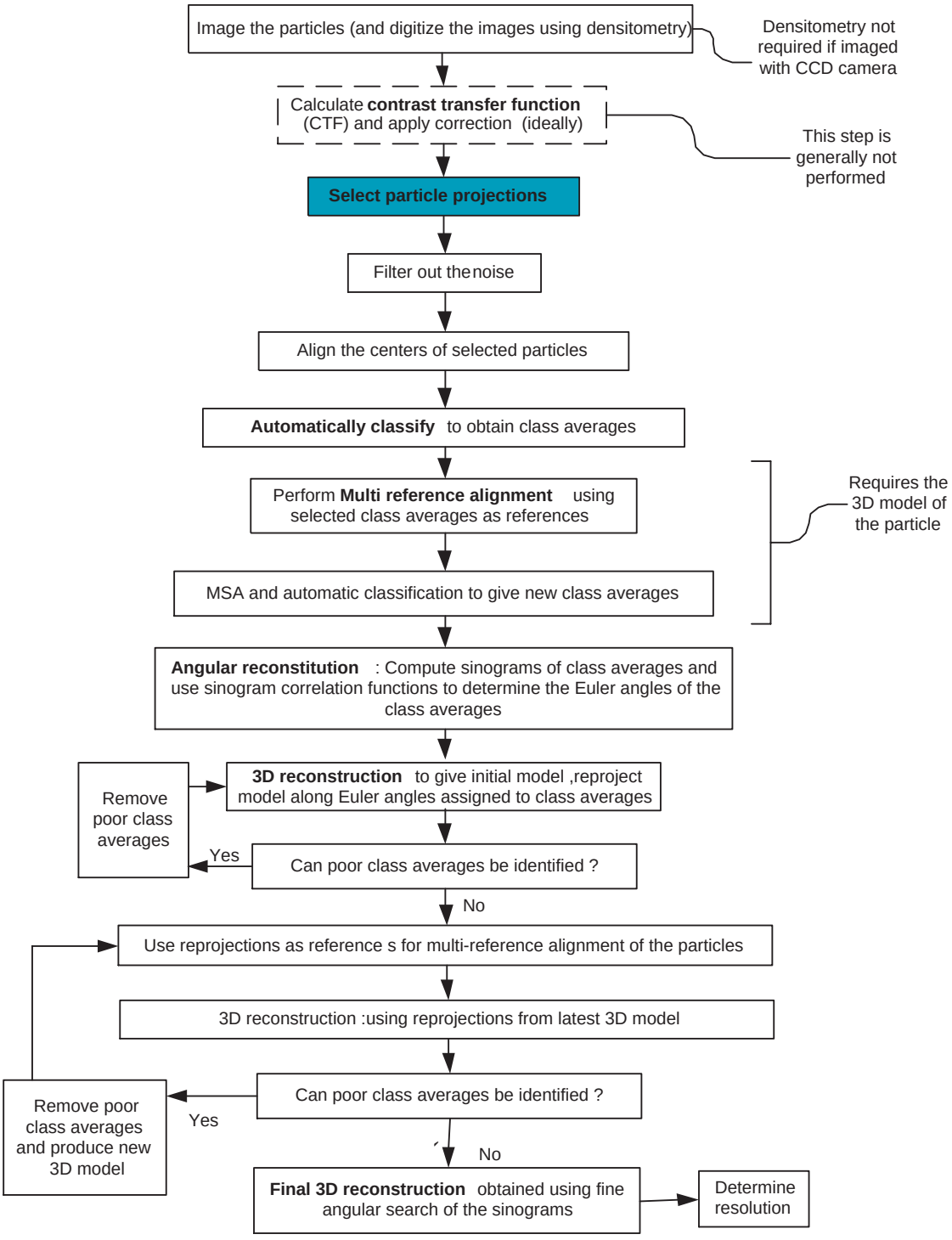


Figure 1.1: A schematic for the process of 3D reconstruction of macromolecules. The emphasis of this work is the step depicted in the blue box. Adapted from [64].

tures routinely determined by crystallography methods vis-s-vis single particle methods. Thus there is a need to bridge this gap.

As shown in Figure 1.1, the initial step in three-dimensional structural studies of single particles and viruses after electron micrographs have been digitized is the *identification* (boxing) of particles images. Traditionally, this task has been accomplished by manual or semi-automatic procedures. However, the goal of greatly improving the solution of structure determinations to 7 Å or better comes with a requirement to significantly increase the number of images. Though 50 or fewer particle projections often suffice for computing reconstructions of many viruses in the 20 Å - 30 Å range [2], the number of particle projections needed for a reconstruction of a virus with unknown symmetry, at $\sim$ 5 Å may increase by three orders of magnitude. Hence, the manual or semi-automatic particle identification techniques create a burdensome bottleneck in the overall process of three dimensional structure determination [50] [30]. It is simply unfeasible to manually identify tens to hundreds of thousands of particle projections in low contrast micrographs, and, even if feasible, the manual process is prone to errors. In the case of particle identification being included in the feedback loop of three dimensional reconstruction, it would be difficult to include a manual or semi automatic technique as a part of the loop.

At high magnification, noise in cryo-TEM micrographs of unstained, frozen hydrated macromolecules is unavoidable and makes automatic or semi-automatic detection of particle positions a challenging task. Since biological specimens are highly sensitive to the damaging effects of the electron beam used in cryo-TEM, minimal exposure methods must be employed and this results in very noisy, low contrast images.

1.2 Problem Formulation

High noise content, non uniformity in the pixel intensity distribution, and the need for identifying projections of asymmetric particles are the major problems to be addressed by any particle identification algorithm. Given an image $f(x, y)$ the objective is to ascertain the regions in $f(x, y)$ that correspond to particle projection. Since the step "Reference free alignment to center particles" in Figure 1.1 computes the centers of the particles, an accurate centering of the particles is not essential. The objective is to minimize both the false hit and false miss rates though the emphasis is more on the former due to it's effect on the 3D reconstruction.

Due to the extensive computational needs of such algorithms [42], a parallel version of the algorithm would have to be developed. There is currently a lack of standard benchmark for relative performance analysis of particle identification algorithms. A benchmark is essential for comparison of the performance of particle identification algorithms. In addition of the benchmark, performance metrics are required for a quantitative comparison. False hits and false misses are currently used as the metrics. Their direct effect on the 3D reconstruction is not known clearly. The performance of particle identification algorithms should be expressed in a much richer set of metrics. For the projections identified by an automatic identification algorithm, such metrics could be used in selection and rejection of projections during the later stages based on the quality of the projections. It would be possible to use only the certain best projections.

1.3 Organization of the Thesis

This work focuses on automatic methods for identification of biological macromolecular particle projections from electron micrographs. A description of the process of three dimensional reconstruction of biological macromolecules using single particle analysis is given in Chapter 2. This is followed by a survey of current methods for particle identification. Chapter 3 presents the theoretical foundations for the algorithms presented in the later part of the thesis. This includes a description of hidden Markov random fields, expectation maximization, and anisotropic diffusion based filtering. Chapter 4 gives a description of model based method for particle identification where model refers to a very low resolution 3D model of the macromolecule. Boxing strategies are also introduced in Chapter 4 where we describe the techniques developed for identification of the particle projections in the segmented images. We give some performance data of our algorithm on a collection of images from our virtual workbench. A parallel implementation of the algorithms is presented in section 4.4 and issues related to load balancing are discussed. Details of the virtual workbench are presented in Chapter 5. Characterization of micrographs and algorithms is described in Chapter 6. We conclude with presenting the conclusions and a detailed list of directions for future work.

1.4 Research Contributions

A list of the research contributions is given below.

- A new technique for particle identification from electron micrographs has been proposed [72] and examined in the context of single particle analysis of macromolecules. The technique involves a novel approach for segmentation of micrographs after filtering them using non-linear anisotropic diffusion.

- Boxing methods based on prior knowledge about the size and/or a three dimensional model of the macromolecule have been introduced. Parallel implementation of these methods have been examined and issues related to it are discussed.

- A web accessible workbench was developed that is generally accessible to the researchers. The workbench consists of tools for characterization of micrographs and algorithms for particle identification. Performance metrics for particle identification algorithms were developed and incorporated in software tools in order to make quantitative comparison between particle identification algorithms over a class of micrographs.

Publications

1. **Vivek Singh**, Dan C. Marinescu, and Timothy S. Baker, Image segmentation for automatic particle identification in electron micrographs based on hidden Markov random field models and expectation maximization , Journal of Structural Biology, 145(1-2):123–141.

2. **Vivek Singh**, Yongchang Ji, and Dan C. Marinescu, A Parallel Algorithm for Automatic Particle Identification in Micrographs, Lecture Notes in Computer Science 3402:354–367, 2005.

3. Yongchang Ji, Dan C. Marinescu, **Vivek Singh**, G. Marinescu, Computational Aspects of Virus Structure Determination at High Resolution, Handbook of Theoretical and Computational Nanotechnologies, American Scientific Publishers, Stevenson Ranch, Ca., 2004 (to appear)

4. **Vivek Singh**, Yongchang Ji, and Dan C. Marinescu, A Parallel Algorithm for Automatic Particle Identification in Electron Micrographs, VECPAR 2004, Valencia, Spain.

5. **Vivek Singh**, Orit K. Schwartz, and Dan C. Marinescu, A virtual workbench for particle selection algorithms, In preparation.

CHAPTER 2

BACKGROUND AND RELATED WORK

Freedom is not worth having if it does not include the freedom to make mistakes.

— *Mohandas K. Gandhi.*

2.1 Particle Identification in Single Particle Analysis

Crystallographic data is obtained from crystals of the sample under investigation. These crystals are arranged in an order thus boosting the signal over noise. Unfortunately, we are not afforded such luxury in single particle analysis where the particles are randomly embedded in vitreous ice. Subsequent imaging of such randomly oriented particles gives projections along randomly distributed directions. Such lack of knowledge of particle orientation corresponding to each projection in single particle techniques creates problems during the 3D reconstruction. These problems are addressed by computational methods.

2.1.1 Single Particle Analysis

Single particle analysis of biological macromolecules is based on the idea of signal boosting by averaging. Given a collection of closely related responses

$$f_i(x, y), \quad where\ i = 1 \ldots n\ ,$$

that are noisy, noise being additive, the average $\overline{f}$ response has a boost in signal by a factor of $n^{1/2}$ [11]. Such signal boosting by averaging can be seen as a way of trying to achieve the signal boosting by crystallization in crystallographic techniques. Single particle analysis does not require crystals to be prepared as a prerequisite at the cost of loss in resolutions achieved in the `3D` reconstructions. The steps that comprise single particle analysis as illustrated in Figure 1.1 are briefly described below.

2.1.1.1 Specimen Preparation

Prior to imaging the macromolecules, they have to be processed in order to obtain desirable images. The specimen preparation must avoid the collapse of structures since the specimen are view in the vacuum of the electron microscope (EM). Biological samples are sensitive to electron bombardment which leads to ionization where incident electron collide inelastically with the specimen, forming highly reactive ions and free radicals. These may disrupt the bonds in the molecules. For Cryo-EM imaging, the specimen is applied to a "holey" carbon grid which is glow-discharged to maintain it at a slightly hydrophillic level. The grid is then blotted with filter paper and plunged into ethane maintained at liquid nitrogen temperature. The thickness of ice is controlled by varying the blotting time. The

specimen is cryo-transferred to the microscope specimen holder. An interesting development in this area was the development of a technique [5] for time resolved imaging to capture images of macromolecules in their various conformation states. In [5] an atomizer spray method was devised to spraying droplets of acetylcholine along with ferritin, a marker, onto a grid containing ordered membrane arrays of the nicotinic acetylcholine receptor just before it is plunged into the ethane. The marker indicated the regions on the grid where droplets landed, so that only membranes exposed to the spray could be selected for processing. The time resolution of the spray method is a few milliseconds, so that they were able to trap the activated stated of receptor, which has a lifetime of 10ms.

2.1.1.2 Digitization and densitometry

Densitometry converts the information in EM film images into a digital form so that it is suitable for computational processing. As per Shannon's sampling theorem [69], the images must be scanned at a sampling rate that is at least half the desired molecular resolution at the specimen level. However usually the images are digitized at quarter the desired molecular resolution at the specimen level in order to contain the interpolation errors. A line scanner is preferred to point scanner due to the speed of scans.

2.1.1.3 Contrast transfer function

The projections of particles on the micrographs are not the true projections of the macromolecule at a particular direction of view. The EM, distorts the true repre-

sentation by selective filtering of spatial frequencies during the imaging process. The contrast transfer function (CTF) describes the fidelity with which the different spatial frequencies are transmitted by the electron lenses.

The image formation process in the microscope can be modelled as a point spread function h. The cryo image i obtained is a noisy version of the true projection ϕ of the electron density potential function of an object convolved by h.

$$i(\mathbf{r}) = h(\mathbf{r}) \otimes (\phi(\mathbf{r}) + n_b(\mathbf{r})) + n_a(\mathbf{r}), \tag{2.1}$$

where $\mathbf{r}$ is a vector in R^2 representing a real space point, n_b, denotes noise before image formation, and n_a, denotes noise after the image formation. The noise terms $h(\mathbf{r}n_b(\mathbf{r}) + n_a(\mathbf{r})$ are combined and represented as $n(\mathbf{r})$ and $h(\mathbf{r}$ is decomposed into a convolution of the contrast transfer function $c(\mathbf{r}$ and the envelope function $e(\mathbf{r}$. This leads to the following equation –

$$i(\mathbf{r}) = c(\mathbf{r}) \otimes e(\mathbf{r}) \otimes \phi(\mathbf{r}) + n(\mathbf{r}), \tag{2.2}$$

Fourier transformation of expression 2.2 yields

$$I(\omega) = C(\omega) \otimes E(\omega) \otimes \Phi(\omega) + N(\omega), \tag{2.3}$$

where ω is a vector representing the spatial frequency, and C represents the CTF. C is a complicated parametric function, which takes into account the effects of voltage, defocus and spherical aberration of the microscope, among others. Under the weak phase and weak amplitude, the CTF can be expressed as –

$$C(\omega) = \sqrt{1 - C_a^2}\ sin(\chi(\omega)) + C_a\ cos(\chi(\omega)) \tag{2.4}$$

where $\chi(\omega)$ is given as

$$\chi(\omega) = \pi\lambda(\Delta f\omega^2 - \frac{1}{2}C_s\omega^4\lambda^2). \tag{2.5}$$

and C_a is the amplitude contrast, C_s is the spherical aberration constant of the lens and Δf is the defocus. The envelope function can be approximated with the following parametric form called the B-factor [95].

$$E(\omega) = e^{-B\omega^2} \tag{2.6}$$

2.1.1.4 Particle identification

The particle projections present in the digitized micrographs have to be identified for processing in the succeeding steps. The number of particles to be selected for `3D` reconstruction at a specified resolution is a function of the specified resolution, the noise content of the micrographs, and the symmetry of the particles. An approximate location of particle projections in the micrograph generally suffices since the centering done in the following step. A band pass filtering may be done to lower the shot noise (high spatial frequency) and the gradual fluctuation in the average pixel intensities across the image (low spatial frequency). Generally gaussian based filters are used for band pass filtering to avoid introduction of artificial artifacts due to filtering. A detailed review of particle identification methods is given in section 2.1.2.

2.1.1.5 Reference free alignment

The centering of particles is achieved by reference free alignment where the data set of identified particles is compared to a rotationally averaged sum of band passed filtered particles. The alignment is achieved using cross-correlation functions (CCF). The CCF shows a peak at the position where a common motif in two images overlap. The reference free alignment is repeated several times until there is a nominal change in further alignment.

Each of the projected images collected after the particle identification step can be considered to be a rotated, translated, and scaled noisy copy of the projection of a true structure. Let us assume that there are N such images corresponding to a particular direction of projection, the true structure is represented as T and the noise is gaussian (G). Each projection can then be represented as –

$$proj_i = Z_j \;\; T(\theta_i) + G_i \;, \quad i \;=\; 1...N \tag{2.7}$$

where Z_j is the scaling factor for the projection obtained from the j^{th} micrograph, θ_i is the rotational and translational version of the true structure T and G_i is the gaussian noise with parameters $\mu_i,\ \sigma_i$ for the i^{th} projection $proj_i$. The goal of the reference free alignment step is to determine for each projection $proj_i$, the parameters θ_i.

2.1.1.6 Multivariate statistical analysis (MSA)

Multivariate statistical analysis is used for classification of the centered particle images. As a preliminary step, correspondence analysis step is used obtain the

principal components of variation in the set of images. Each image of size $x \times y$ is represented as a point in a xy dimensional space. A χ^2 metric is used as the distance between two images in this high dimensional space. The original high dimensional space is converted into a lower dimensional space with the top few (typically 5-10) eigenvectors as principal axes.

2.1.1.7 Classification and multireference alignment

For a specified number of classes automatic classification in the lower dimensional space mentioned above (section 2.1.1.6) is done using unsupervised classification algorithms such as k-means classification. Class averages are computed for each class in order to boost the signal to noise ratio (SNR). Each class is further processed for alignment within the class using the same procedure as section 2.1.1.6. Outliers may be rejected in this step to obtain a better SNR.

2.1.1.8 Angular Reconstitution

The views for the classes of projection images are obtained using angular reconstitution which is based on the common line projection theorem [81]. It states that two `2D` projections are of the same `3D` object have at least one `1D` line projection in common. In order to find this common line projection between the class averages of two classes, the sinogram of each class average is computed. Sinogram correlation functions (SCF) are computed for each pair of class averages by computing the correlation between each of the lines of the two sinograms.

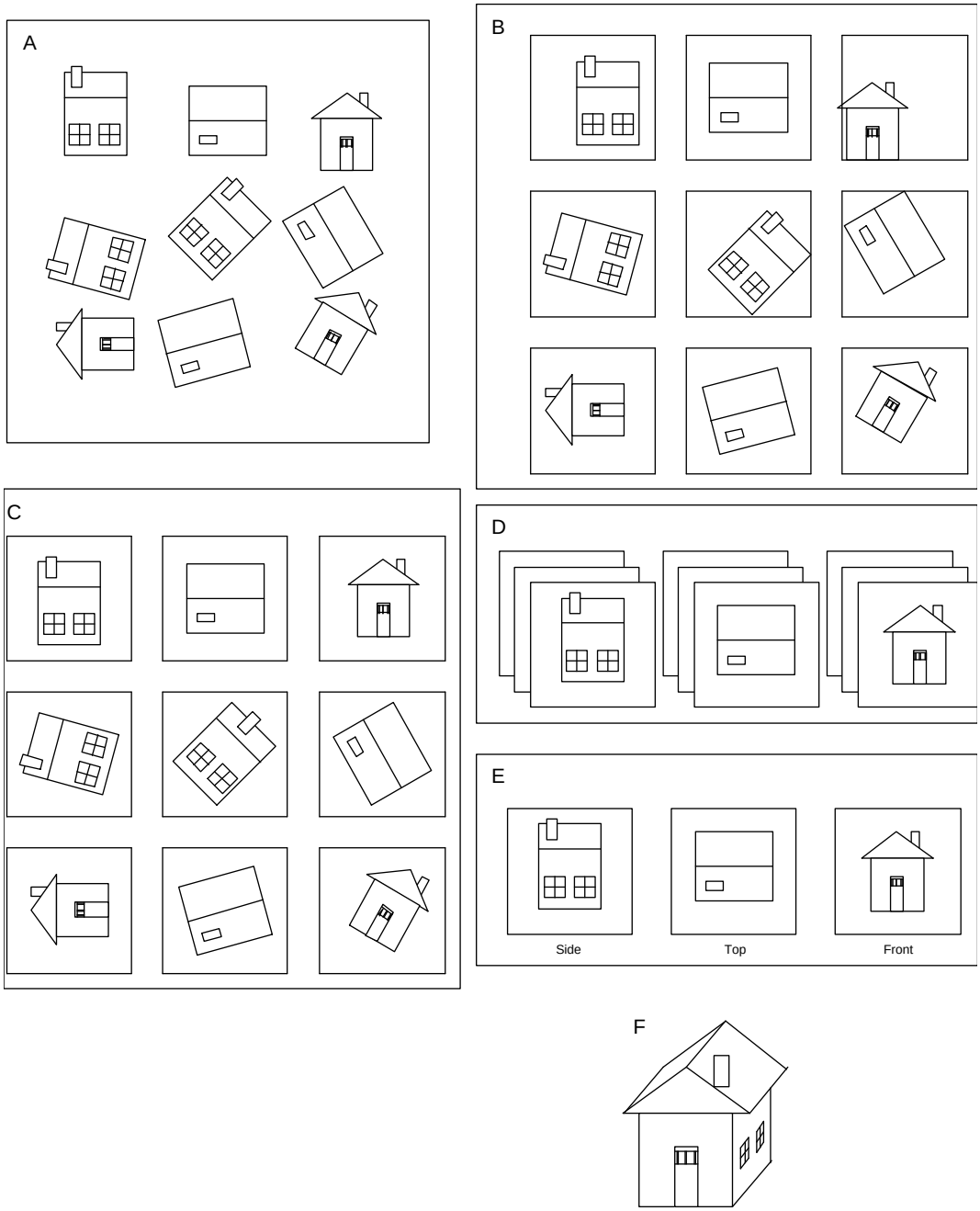


Figure 2.1: An illustration three dimensional reconstructions in terms of image processing. Adapted from [78].

For the sake of illustration, let us assume that we have a class averaged projection image $class_A$ obtained from the experiment and we have two projections obtained from the tentative 3D model $proj_B$ and $proj_C$ with their corresponding sinogram functions as $sino_A$, $sino_B$, and $sino_C$. The sinogram correlation functions are obtained for each pair of the class averages by computing line by line correlation of the corresponding sinograms i.e. $SCF_{A,B}$ is calculated by correlating sinograms $sino_A$ and $sino_B$, and $SCF_{A,C}$ by correlating sinograms $sino_A$ and $sino_C$. We can then assign euler angles to the class average A by finding the global maximum in all its sinogram correlation functions $SCF_{A,B}$ and $SCF_{A,C}$, with respect to projections B and C of the 3D model.

2.1.1.9 3D Reconstruction

The initial 3D reconstruction is obtained by back projection of the class averages along their assigned Euler angles [28] with the use of the projection theorem [12]. The projection theorem is illustrated in 2.2.The 3D reconstruction is then reprojected along the Euler angles assigned to the class averages. Poor class averages are removed from the dataset. The remaining class averages are used to create the 3D reconstruction again. The 3D reconstruction is iteratively refined by using the reconstructed model obtained in the preceding steps of the refinement. The resolution of the 3D reconstruction is evaluated by measuring the Fourier shell correlation (FSC) between two independent 3D reconstructions each based on half of the class averages[83].

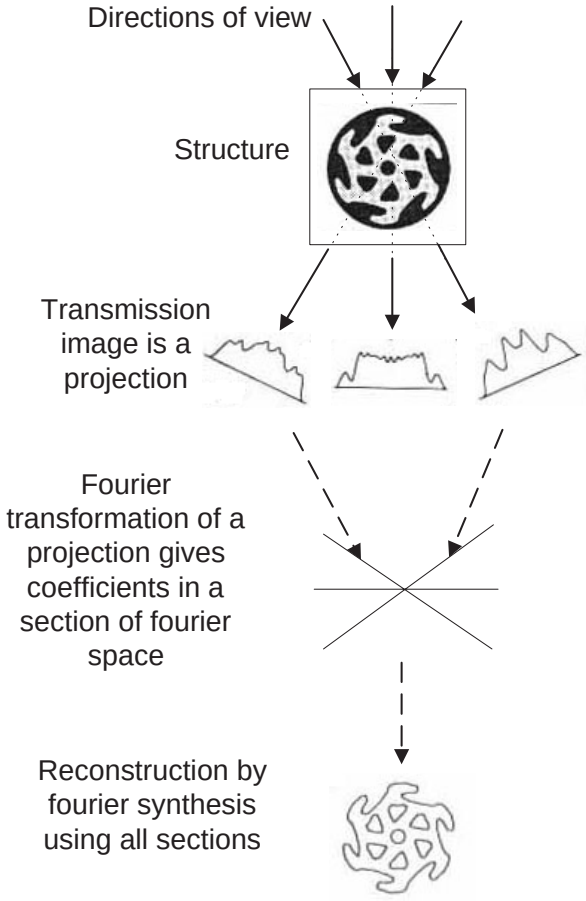


Figure 2.2: An illustration of the projection theorem. Adapted from [12].

2.1.2 Particle identification Methods

Over the past two decade the use of Electron Microscopy (EM) for studying structures of large biological macromolecules has seen tremendous advancement as evidenced by ever increasing resolution of 3D reconstructions of macromolecules. Such successful pursuits have utilized tens of thousands of images of particle projections to achieve higher resolutions. Particle identification is the first step in processing of the micrograph images. Significant user interaction is required in identification of particle projections which are done either manually using interactive graphics software or using computer assisted semi-automated methods. Some of the commonly used methods for particle identification [49] are reviewed below.

2.1.2.1 Template Matching Based Methods

In a typical template matching based method, a reference image, serving as a template, is evaluated for a match within a given micrograph image for each possible location of the reference over the micrograph. As illustrated in Fig.2.3, $f(x, y)$ i.e. image A and $g(x, y)$ i.e. image B may, without loss of generality, be assumed to be the micrograph and reference image respectively. As shown the two images are assumed to be of the same size. The fact that the reference image is almost always much smaller than the micrograph can easily be accommodated by assuming that the support of $g(x, y)$ extends over the size of support of $f(x, y)$ with the value of $g(x, y)$ beyond it's original support being 0. With the aforementioned assumptions, the correlation map of image A and image B is given by the cross correlation function $c(x', y')$ defined as the following –

$$c(x', y') = \sum_x \sum_y f(x, y) g(x + x', y + y') \tag{2.8}$$

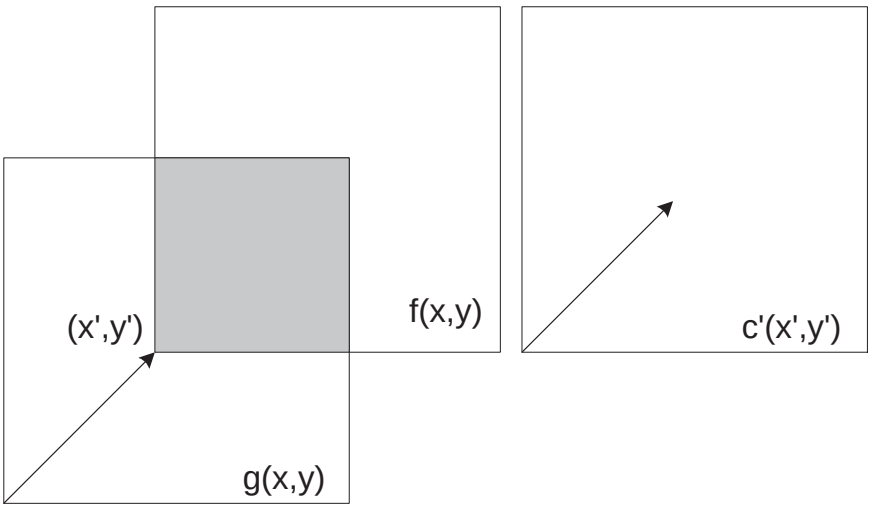


Figure 2.3: An illustration of cross correlation. Scalar product of image A and image B over the grey region corresponds to c(x',y'). Adapted from [20].

The cross correlation map, which is the ordered set of values of the cross correlation function over the support of the given micrograph, can be efficiently computed by using the convolution theorem from Fourier transform theory. Following the theorem, the cross correlation function can be written as

$$c(x', y') = F^{-1}\{F\{f(x, y)\}F\{g(x, y)\}^*\} \tag{2.9}$$

where F indicates the Fourier transform operation, F^{-1}, the inverse Fourier transform operation and $F\{g(x, y)\}^*$, the conjugation operation on Fourier transform of $g(x, y)$. As can clearly seen, the term $F\{g(x, y)\}^*$ acts as a filter modifying the Fourier transform of the original signal before transforming the product back to the real space. In the reciprocal space i.e. the frequency space, $F\{g(x, y)\}^*$,

a matched filter, may be represented as $H(s_x, s_y)$. The matching locations of $g(x, y)$ in the image $f(x, y)$ can be obtained by locating the peaks in the response of image $f(x, y)$ after it is filtered by a matched filter correspond to $g(x, y)$. However, presence of noise in general, and in the case of micrographs shot noise and noise in the form of non uniform variation of image intensity, can make peak detection a difficult task. Another problem associated with particle identification using cross correlation based methods is their high sensitivity to variations in the projections due to rotations in the image plane. The alternative is to use a gaussian function as a template. The problem of non uniform variation is overcome using correlation coefficient instead of cross correlation function.

2.1.2.2 Low Level Feature Based Methods

Low level feature based methods are essentially methods that operate on low level cues obtained by operations such as edge detection and image segmentation i.e. pixel classification. An advantage that these methods, especially the edge based methods, command over template based approaches is that they are immune to noise in the form of non uniform variation in image intensity. This can be attributed to their use of local neighborhood based computation for obtaining the features. However the overwhelming presence of shot noise plagues these methods too. Once the low level features are obtained, such methods are faced with the problem of identifying the representations of objets i.e. particle projections from a collection of features distributed over the support of the micrograph.

Edge detection Based Approaches An edge in an image is a boundary or contour at which a significant change occurs in some physical aspect of an

image, such as the surface reflectance, illumination, etc. [38]. For an image function, edges can be defined as zero-crossings of the Laplacian or the maxima of the gradient modulus in the gradient direction [9]. The edge detector proposed by Canny [9] is the most commonly used edge detector. The image is initially smoothed by *Gaussian convolution with a kernel variance of* σ. Then, a simple `2D` first derivative operator is applied to the smoothed image to highlight regions of the image with high-valued first spatial derivatives. Edges give rise to ridges in the gradient magnitude image. A process known as *non-maximal suppression* is then applied wherein the algorithm tracks along the top of these ridges and sets to zero all pixels that are not actually on the ridge top, so as to give a thin line in the output. The tracking process is controlled by two thresholds, $High > Low$. Tracking can only begin at a point on a ridge higher than $High$. Tracking then continues in both directions out from that point until the height of the ridge falls below Low. This helps to ensure that noisy edges are not broken up into multiple edge fragments. The effect of applying a Canny edge detector with parameters $Low = 0.03$, $High = 0.07$, and $\sigma = 3$ to the image of frozen-hydrated virus particles is illustrated in Figure 2.4. The parameters of the Canny edge detector are selected on an ad hoc basis. Although the optimal values of parameters required by the edge detection algorithm may not vary much from one micrograph to another when the micrographs were collected under similar experimental conditions, nevertheless the ad hoc selection of the parameters is not very convenient. To address this issue, automatic selection of parameters for the Canny edge detector have been developed [89]. However single scale approach limits the quality of edges detected in the presence of high noise.

The low SNR of micrographs necessitates a pre-processing of the micrographs prior to edge detection. Although gaussian filtering is commonly used, the isotropic property of the filter leads to blurring of the particle projection bound-

Figure 2.4: The use of Canny edge detector. a) Small field of view from an electron microscope of frozen hydrated virus particles. b) Same as (a) after application of Canny edge detector. c) The histogram of the pixel intensities in (a) illustrates the narrow dynamic range of pixel intensities.

aries. In order to preserve the edges, non linear anisotropic diffusion and bilateral filters provide an anisotropic smoothing [86] which smoothes the quasi-homogeneous regions more than the regions of discontinuity in the image function.

The approach taken by [27] involves the initial step of detecting edges which is followed by connected component labeling and symbolic processing. The technique proposed in [94] also requires an edge map of the micrograph initially. The *Hough transform* is used to spatially cluster the edges to represent particles. The Hough transform is based upon a voting algorithm [26]. A generalized Hough transform is used for particle projections with irregular geometric shapes. Since both the techniques are based on edge detection, their performance is closely dependent on the performance of the edge detection algorithm. Figure 2.5 shows the schematic for the particle identification algorithm of [93]. A similar approach proposed in [94] has been successfully used for particle identification from micrographs of helical objects.

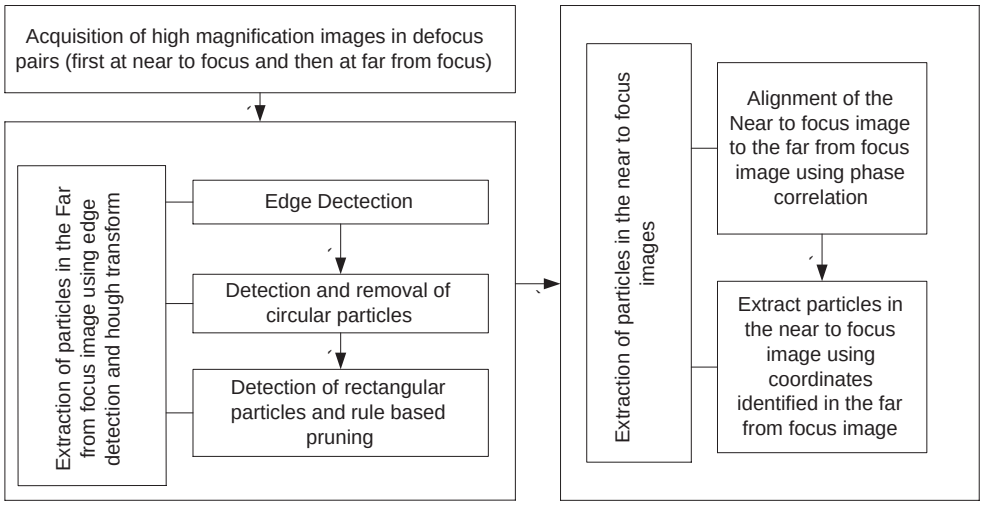


Figure 2.5: A schematic for algorithm in [94].

The Signal to Noise Ratio (SNR), and in general the characterization of the noise in a micrograph, are very important to determine the best technique for automatic particle identification to be used for that micrograph. Noise estimation could help the automatic selection of the parameters of an edge detection algorithm.

Elder and Zucker [14] describe a method of automatic selection of the reliable scales for the computation of the second order derivative of the intensity in the image field. A scale is assigned to each pixel for the computation of the second order spatial derivative. The noise is assumed to be *additive white Gaussian*. The idea is to select scales for each pixel that, given the magnitude of the standard deviation of the noise, are large enough to provide a reliable derivative estimate. If the second order derivative estimate at a pixel for a particular scale is below a threshold which is governed by the standard deviation of the noise, then the scale is increased until the second order derivative at such a scale exceeds the threshold. The only parameter required by this method is the standard deviation of the noise content. Thus, a procedure for the estimation of the standard deviation of the

noise in micrographs would lead to an automatic edge detection algorithm. Lee and Hoppel [31] describe a fast method to compute the standard deviation of the noise in an image. This method is based upon two assumptions (a) identically distributed (i.i.d) noise, and (b) a linear noise model. This means zero mean additive noise and unit mean multiplicative independent noise. Consider three random variables representing $Z_{(a,b)}$ - the observed intensity at the pixel with coordinates (a, b), $X_{(a,b)}$ - the noise free true intensity at the pixel with coordinates (a, b), W - a random variable that represents a zero mean additive i.i.d. noise with a standard deviation s_w. V - a random variable that represents a unit mean multiplicative noise with standard deviation s_v. Then,

$$Z_{(a,b)} = X_{(a,b)}V + W \tag{2.10}$$

For a homogeneous block

$$\bar{Z} = \bar{X} \tag{2.11}$$

$$\bar{W} = 0 \tag{2.12}$$

$$\bar{V} = 1 \tag{2.13}$$

thus

$$var(Z) = s_v^2 * \bar{Z}^2 + s_w^2. \tag{2.14}$$

Thus there is a linear relation between $var(Z)$ and $\bar{Z}^2$ with the coefficients being the variances of the respective linear noise types – additive with s_w and multiplicative with s_v. The algorithm to estimate s_w and s_v consists of the following steps

1. Divide the image into blocks of a given size, e.g., 4x4 or 8x8 pixels.

2. Compute the variance (i.e., $var(Z)$) and square of the means (i.e., $\bar{Z}^2$) of the pixel intensities in each block.

3. Use a least squares method to approximate a linear relation between the variances and the squares of the mean. A least square solution with negligible slope is indicative of an additive noise.

2.1.2.3 Other Approaches

The method described in [43] identifies particle projections that are circular in shape. The micrograph image is initially histogram equalized and sub-sampled. Subsequently pixels belonging to the same object are identified by a double scan technique. A schematic of the procedure is shown in Figure 2.6. A method for identification of projections that are ring shaped is described in [35], which is essentially a convolution of the normalized positive ring with a negative circle placed in the center of the ring [24].

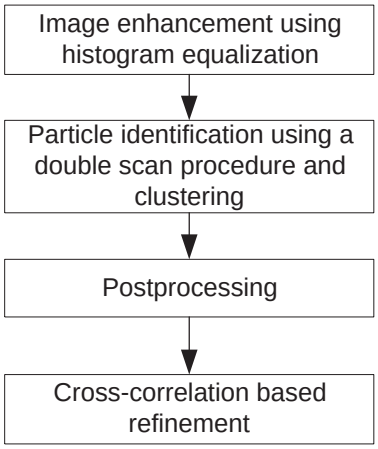


Figure 2.6: A schematic for the crosspoint method proposed in [43].

The method proposed in [59] uses Fischer discriminant analysis on textures for identification. The user is initially required to train the algorithm with a few truth samples. Based on nine features extracted from these samples, a Fischer discriminant classifier is generated and is further used to identify regions in the image that are similar to the training classes.

Recently, neural network methods have been applied to particle identification [44]. Neural networks have the ability to take into consideration a great degree of complexity for identifying projections of single particles. A schematic of the particle identifications process using the neural networks is shown in Figure 2.7.

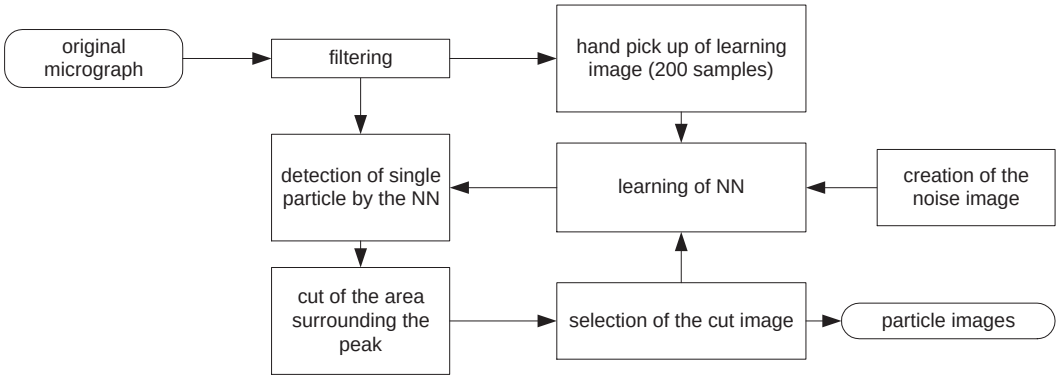


Figure 2.7: A schematic for the neural network based particle pickup method proposed in [52].

2.1.3 Caveats

The goal of the single particle analysis is to obtain a faithful 3D representation of macromolecular particles. The effect of identification of particles on 3D reconstruction is not completely understood. Hence, the performance of particle identification algorithms can not be merely represented by simple metrics such

as the false hit and false miss rates. The quality of the identified particles need to be assessed alongside their effect on the final 3D reconstruction. This entails the development of metrics for the quality of identified particles, techniques for assessment of the effect that an ensemble of identified particles corresponding to a projection have on the improvement in 3D reconstruction, and development of programs for 3D reconstruction of macromolecules based on "good" projections of the particles, where "good" is representative of the quality of the identified projections. As an after thought, single particle techniques for 3D reconstruction may be improved by integrating the 3D structures of the constituents of these macromolecules obtained using such techniques as X-ray crystallography and NMR spectroscopy during the reconstruction process.

The performance of particle identification algorithms is heavily dependent on the quality of micrographs and the geometry of the particle projections present in the micrographs. However due to the lack of a standardized collection of dataset, the comparison among the particle identification algorithms based on their performance is not possible, a need that we try to a address with the workbench as discussed in chapter 5 and chapter 6. Table 2.1 depicts the performance of a few particle identification algorithms. However, since the performance results are not available on the same set of data, a direct comparison is not possible. Moreover, the metrics 'false positives' and 'false negatives' are not sufficient enough to judge the performance of the particle identification methods. More rigorous metrics must be required. Some such metrics are discussed in section 6.2.1.

Table 2.1: A performance comparison of the particle identification algorithms is given with the data obtained from the respective publications as indicated.

Sl. no.	Publication reference	Name	False (%) Positives	False (%) Negatives
1	[43]	crosspoint	4	11
2	[94]	Zhu et. al.	21	12
3	[72]	HMRF-EM	19	9
4	[39]	EMAN	23.7	43.4
5	[61]	FindEM	16.6	2.4
6	[71]	Sigworth	4.5	23.2

2.2 Semi-automated systems for single particle analysis

Over the past few years, some automated system for single particle analysis have been developed. However, none of the systems address the issue of comparison between several particle identification algorithms. Indeed the aim of all of them is to further the automation of the process of three dimensional reconstruction of macromolecules rather than benchmarking and comparison of algorithms. The automation systems can be broadly classified as those that aim toward automating the data acquisition process and those that aim toward automating the analysis of the data once the data is available. One such suite is Leginon ([10]) which is currently the most comprehensive of the suites available for automated data collection for single particle analysis. Although, the primary aim of Leginon is to automate the data acquisition process, it also features some rudimentary semi-automated data analysis tools too. Leginon is a fairly general and extensible suite for image image acquisition and data collection for automated microscopy. The system consists of reusable modules called *nodes*. The nodes

may be connected to create an *application*. A set of connected nodes communicate by generating *events*. Applications are designed by selecting the nodes required for the application, the order of execution of the nodes and the event semantics. Graphical tools simplify the task of creating new applications. Once an application has been generated, it can be stored in the persistent storage and reused later. A logging facility is provided to allow monitoring of progress. A relational database is used to store data between runs. The application can be distributed over several machines due to the availability of executable codes for nodes on several types of machine configurations. The Figure 2.8 illustrates the architecture of a Leginon application.

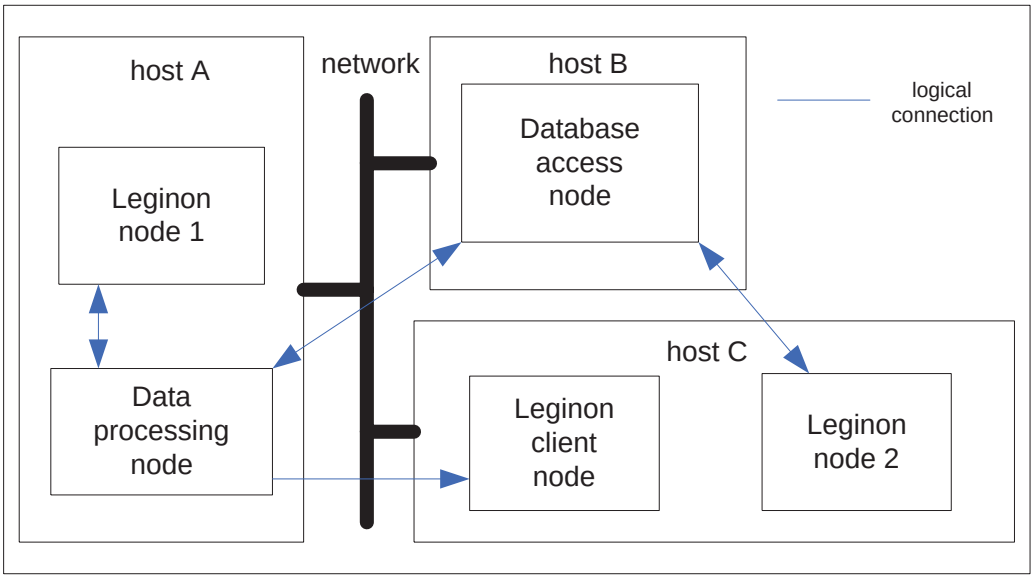


Figure 2.8: A typical Leginon application which is constructed using pre-defined modules called *nodes*. The logical connections shown in blue are called *events*. The physical layout of the application is also shown. Adapted from [77]

Another system quite widely used in single particle analysis is EMAN [39]. In contrast to Leginon, EMAN (like Imagic [84], Spider [21] and XMIPP [76]) is

a system for semi-automated data analysis. It is a collection of programs mainly written in $C++$. The process of three dimensional reconstruction is achieved using this collection of programs as illustrated in Figure 2.9. Particle identification in EMAN is semi automatic and is done by a program called *boxer*. The program is based on the concept of cross correlation (section 2.1.2.1). Traditionally cross correlation methods have been based on generating a rotationally averaged reference image, generating a map (called cross correlation map) by cross correlating the reference image with the micrograph. Boxer uses multiple references due to the variation among the various projections for the same particle. A cross correlation map is generated for which the value at each pixel is the maximum cross correlation for any of the templates. In order to reduce the chance of duplicate selections, low pass filtering is used. A peak searching algorithm extracts the location of all recognized particles. The threshold for peak selection is controlled by the user.

In addition to EMAN, Spider is also a system for semi-automated data analysis. Just as EMAN, the spider system is also a collection of programs for single particle analysis. It includes two methods for particle identification. The first method is based on [59] and is a supervised learning based algorithm where the user selects regions of the micrograph that are 'projections', 'noise', and 'junk'. The algorithm then computes a discriminant function based on the characteristics of these regions as per some statistically computed features. The particle identification for the rest of the micrographs is done using this discriminant function. The second method is based on the local correlation function as described in [60]. The local correlation function is a modification of the normalized correlation function. The performance results of [60] is given in table 2.1.

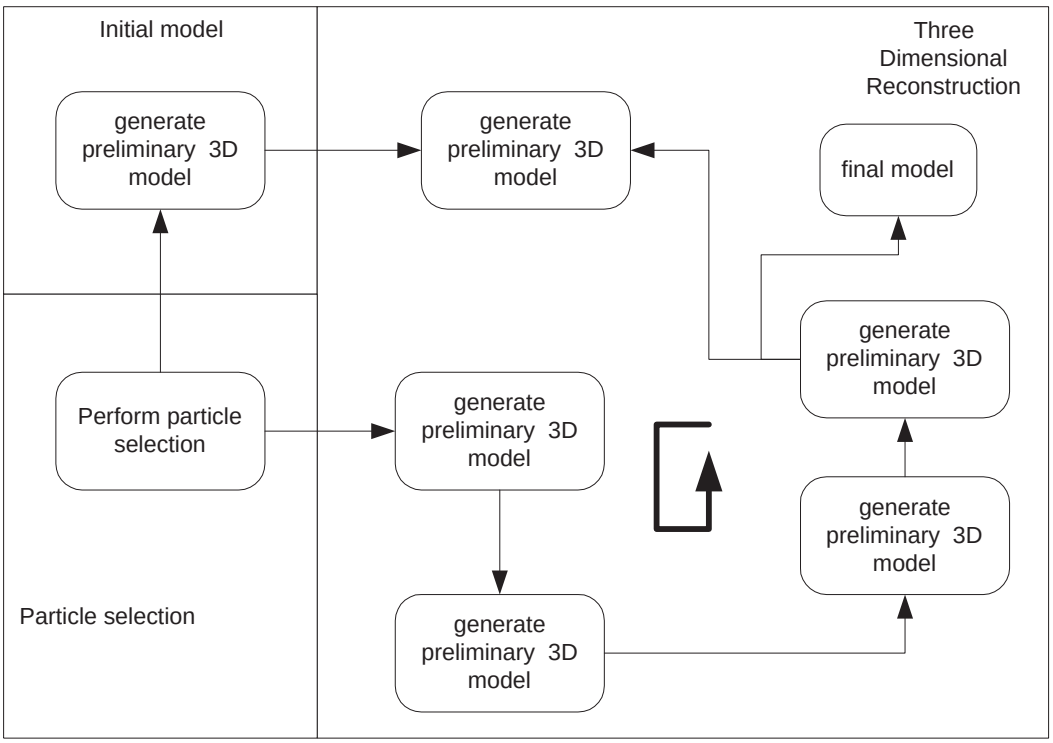


Figure 2.9: The process for three dimensional reconstruction in EMAN is shown here. Adapted from [39]

CHAPTER 3

THEORETICAL FOUNDATIONS

I know nothing except the fact of my ignorance.

— *Socrates.*

The algorithms described in Chapter 4 are based on image segmentation [26]. Image segmentation can be thought of as a labelling problem where the task is to label the image pixels as belonging to one of several groups. Particle identification amounts to a binary filtering operation, in that pixels are marked according to whether they belong to the particle or the background. Hence, such a segmentation could be thought of as labelling of the image field with labels picked up from a binary set {0,1}. One of the popular methods of image segmentation is based on modelling the image as a Markov random field. The realization of the field is presumed to the true image. Here we provide a description of the theory of Markov random fields.

The micrograph images have very low signal to noise ratio (SNR). Hence before segmentation, the images have to be filtered to reduce the noise content. Anisotropic diffusion, a novel technique developed by [57] is used to filter the micrographs prior to segmentation. A description of this technique is presented.

3.1 Markov Random Field Models

The distribution of intensities in an image exhibits a coherency in space. For instance, in an image of a car, a pixel belonging to the car is highly likely to be next to other pixels belonging to the car. Such contextual constraints restrict the solution space for segmentation of an image which is to say that an image with a random distribution of labels is unlikely to be the true segmentation of the image. It is commonly accepted that the pixel intensities in a micrograph exhibit high spatial statistical interdependence, i.e., background pixels have a high probability of occurring next to other background pixels. Likewise, "particle" pixels generally lie adjacent to other "particle" pixels. Markov random field (MRF) theory provides a basis for modelling such contextual constraints. Further, it allows us to formulate particle identification as an optimization problem. A wealth of research done in the field of optimization has been extended to solve the problem of image segmentation through the use of Markov Random field models. We begin with a description of the notations used in the theory of Markov random fields (MRF) and give an exposition of some basic assumptions made. We then give a definition of the MRF, provide a description of the equivalent Gibbs random field. Potential functions, the core of MRF models are presented here in. The specific model of the micrograph image is presented in the next section. However, the segmented image and parameters of the image model are not available initially. Expectation maximization, a technique to estimate parameters for incomplete data, is shown next along with it's use for our specific problem. The high noise content of the images require that they be preprocessed before any analysis. Anisotropic diffusion, one such method for reduction of noise is presented next.

3.1.1 Notations and Basic Assumptions

The task of particle identification may be abstracted as a labeling problem where the aim of the image processing entity is to ascertain for each pixel whether it belongs to the particle projection or the background i.e. assigning a label to each pixel where the label identifies the pixel as either a part of the projection or not a part of it i.e. background. Before we develop the algorithm, we must look at some notations and definitions, and explore some basic assumptions. We use the following notations [7].

$L = \{0, 1\}$ - the set of labels. The elements of the set L are assigned to pixels in the segmented image. The segmented image is represented as a realization of field X.

$D = \{1, 2, \ldots d\}$ - the set of quantized intensities in the observation field, i.e., the micrograph. The intensities of the pixels in the original micrograph are elements of the set D. The original micrograph image is represented as realization of field Y.

$S = \{1, 2, \ldots M\}$ - the set of indices. This set consists of elements that are used to index sites in both X and Y fields.

$R = \{r_i, i \in S\}$ a family of random variables indexed by S. The fields X and Y are two types of R with different constraints.

r - a realization of R. The fields X and Y have realizations x and y respectively.

As mentioned above, X and Y are random fields, where Y represents the observed field i.e., the micrograph, and X represents the segmented image(Figure

[7] *Some of the other notations are developed as they are encountered.*

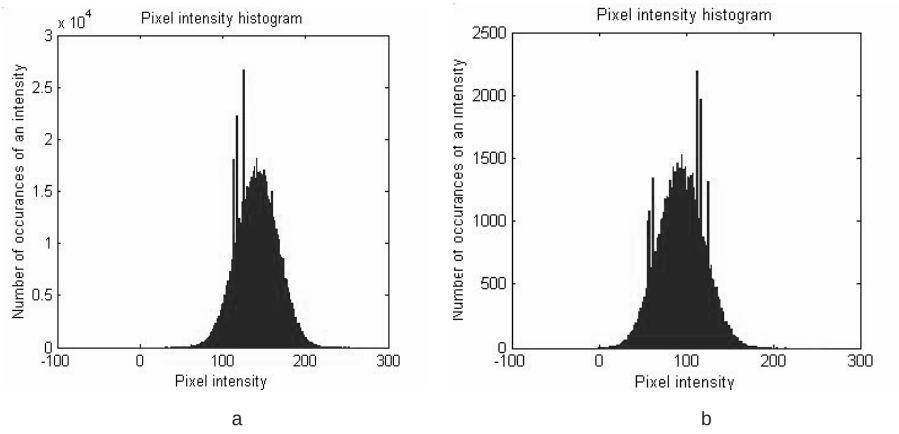


Figure 3.1: The assumption of Gaussian distribution of pixel intensities is supported by the experimental evidence from the histograms of pixel intensities belonging to the background and to particle projections. a) The distribution of the background pixel intensities. All the pixels constituting the background should ideally be assigned the same label. b) The distribution of the pixel intensities inside the projections of virus particles. All the pixels constituting one projection should ideally be assigned the same label, different from the one assigned to pixels from the background.

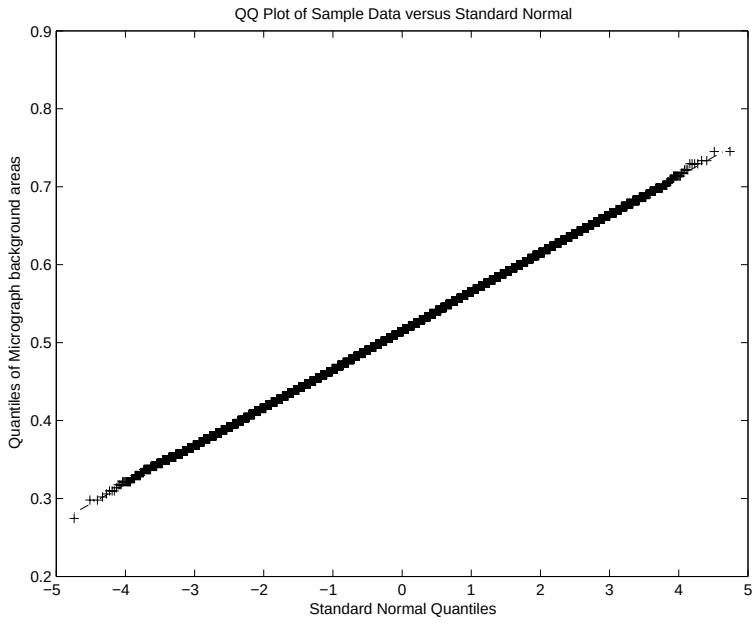


Figure 3.2: The assumption of gaussian distribution of noise in the pixel intensities is supported by the qqplot of the pixel intensities of the background of one of the micrographs

3.5). A *realization of a random field* is a set of values for all the elements of the field. For example, the realization of the field X consists of the set of labels (0 or 1) for every pixel. A realization of the field Y is the set of intensities for each pixel. We denote by x and y a particular realization of the two respective fields.

Let $\mathcal{X}$ be the set of all realizations of the random field X, and similarly let $\mathcal{Y}$ be the set of all realizations of the random field Y.

$$\mathcal{X} = \{x = (x_1, ..., x_M) | x_i \in L, i \in S\}$$

and

$$\mathcal{Y} = \{y = (y_1, ..., y_M) | y_i \in D, i \in S\}$$

The label assigned to the random variable x_i determines the parameters of the distribution for the observed random variable y_i

$$P(y_i | x_i = \ell) = f(y_i; \theta_l), \ \forall \ell \in L \tag{3.1}$$

The function f is assumed to be normally distributed with mean μ_ℓ and standard deviation σ_ℓ, i.e.,

$$f(y_i | x_i = \ell) = \frac{1}{\sqrt{2\pi\sigma_\ell^2}} \ exp \ \{-\frac{(y - \mu_\ell)^2}{2\sigma_\ell^2}\} \tag{3.2}$$

The assumption of a Gaussian functional form of distribution of intensities conditioned on pixel labels is based on the observations presented in Figure 3.1. A histogram of intensities for portions of the background is shown in Figure 3.1(a).

Since MRF is used to model the spatial coherency of intensity distribution in an image, it must have a representation for intensity distribution in a locality and also a metric for the extent of coherence. The coherence is represented using

neighborhood systems and potential functions are used as a metric for the extent of coherence. Potential functions are presented in Section 3.1.2.1.

In an MRF, the sites in S are related to each other via a *neighborhood system*

$$N = \{N_i, i \in S\}$$

where N_i is the set of neighbors of i

$$N_i = \{j \in S : d(i,j)^2 \leq r, j \neq i\}$$

where $d(i,j)$ is the distance between two sites. Note that a site is not a neighbor of itself. An example of a second order neighborhood i.e., a neighborhood in which vertically, laterally, and diagonally adjacent sites are mutual neighbors, is presented in Figure 3.3. The pixel marked x represents the center relative to which the neighborhood is defined.

Closely tied with the concept of neighborhoods is the concept of *Cliques*. The concept of *cliques* comes naturally in order to model the fact that the dependent of a site on it's immediate neighbor is more than it's dependence on a neighbor that is farther hence providing a greater consideration to closer neighbors when estimating the label at a site. A *clique* is a subset of sites in S. $c \subseteq S$ is a clique if every pair of distinct sites in c are neighbors. Single-site, pair-of-sites, triplets-of-sites cliques, and so on, can be defined, depending upon the order of the neighborhood.

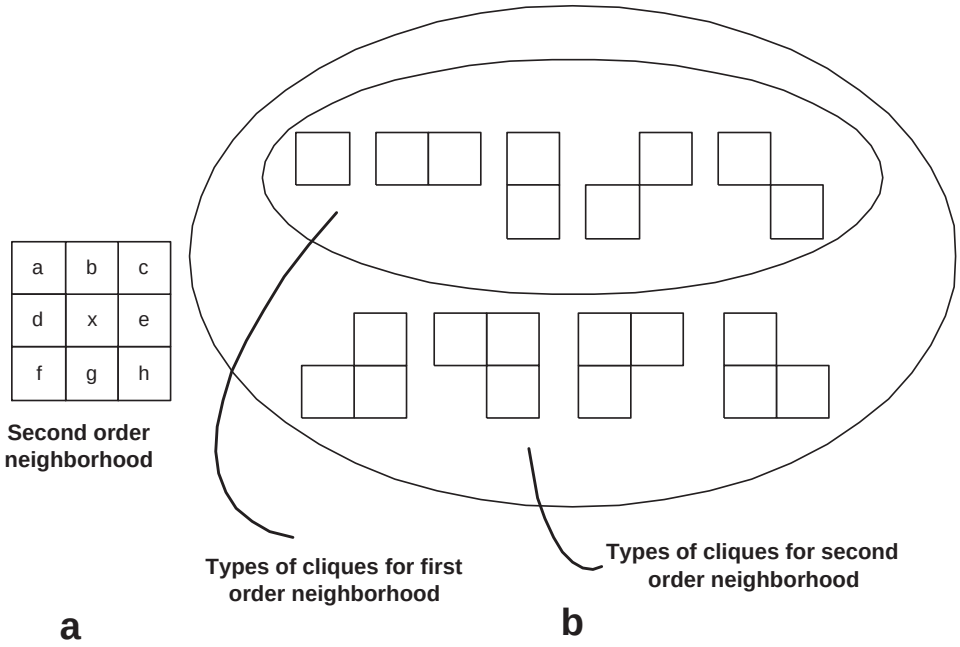


Figure 3.3: (a)First order neighborhood of x consists of {b, d, e, g} and second order neighborhood consists of the set {a, b, c, d, e, f, g, h}. (b) The types of cliques for first order and second order neighborhood.

3.1.2 MRF Definition and Description

A random field X is said to be an MRF on S with respect to a neighbor system N if and only if

$$P(x) > 0 \qquad \forall x \in X$$

and

$$P\left(x_i | x_{S-\{i\}}\right) = P\left(x_i | x_{N_i}\right).$$

The *local characterization* of the MRF defined above simply states that the probability that site i is assigned label x_i can be computed with the knowledge of only the neighborhood of i. Any information more than the configuration of the neighborhood is superfluous. The probability distribution $P(x)$ is uniquely determined by the conditional probabilities. However, it is computationally very difficult to determine these characteristics in practice. We witness a computational explosion as the neighborhood size is even moderately increased.

In order to compute the probabilities of occurrence of a realization of the MRF, we need a way to link the local properties of the MRF with its global properties. The Hammersley-Clifford theorem provides us with such a means. The Hammersley-Clifford theorem establishes a relation between the MRF and the Gibbs distribution. The *Gibbs Distribution* relative to neighborhood system N has a probability measure given by

$$P(x) = Z^{-1} \times e^{-\frac{U(x)}{T}} \quad , \tag{3.3}$$

where Z is the normalizing constant or *partition function* given by

$$Z = \sum_{\forall x_i \in L, \forall i \in S} e^{-\frac{U(x)}{T}}, \tag{3.4}$$

T is a constant called the temperature and U is the energy function

$$U(x) = \sum_{c \in C} V_c(x), \tag{3.5}$$

given by the sum of clique potentials over all possible cliques. C denotes the set of all possible cliques given a neighborhood. The set C consists of all the cliques corresponding to each site in the label field. It may be recalled that the cliques have been defined as a part of Markov Random fields for incorporating interaction between neighbors. As we shall see later, this may be used to ensure a smoothness in the variation of the labels.

Now we need to compute the probability of a particular realization x of X. As mentioned earlier, the Hammersley-Clifford theorem ensures us such a means. It states that X is a MRF on S with respect to a neighbor system N if and only if X is a Gibbs random field on S with respect to the neighbor system N [23]. This provides a simple way of specifying a joint probability for a realization with just the knowledge of conditional probability for a neighborhood at the points. The potential function $V_c(x)$ mentioned in equation 3.5 can take on many forms depending on the application.

3.1.2.1 Potential functions

First order neighborhood is the smallest size neighborhood to convey contextual information. However due to the computational costs associated with higher

order neighborhoods, higher that second order neighborhoods are rarely used. We will consider only the first order neighborhood for description of potential function models. For first order neighborhood, the types of cliques are illustrated in Figure 3.3. For such cliques, the energy function is given by

$$U(x) \;=\; \sum_{\{i\}\in C_1} V_1(x_i) + \sum_{\{i,i'\}\in C_2} V_2(x_i, x_{i'}). \tag{3.6}$$

The most commonly used model are the so called *auto models*, where the potential functions are given by

$$V_1(x_i) = x_i \; G_i(x_i) \;\; \textit{and} \;\; V_2(x_i, x_{i'}) = \beta_{i,i'} \; x_i \; x_{i'} \tag{3.7}$$

where $G_i(\cdot)$ are arbitrary functions and $\beta_{i,i'}$ are constants for interaction between pair of sites i and i'. An auto-model is said to be an *auto-logistic* model, if the x_i's take on values in the discrete label set $\{0,1\}$. As we will see later, the objective is to minimize the sum of energies of all the elements of the set of cliques, ensuring that X_i, a particular realization of the field of labels X, is smooth. But the field X representing the segmented image must be faithful to the measured data (the micrograph). Hence, we define a model of the image that should be integrated with the MRF model.

3.1.3 The Image Model

There are two random fields involved in the model of the image. The random field Y represents the observed image. Y is a random field that does not exhibit any neighborhood relationships, i.e., the random variables y_i are independently distributed. The micrograph is a realization y of Y.

The random field X represents the segmented image and is assumed to exhibit MRF properties. The state of X is not observable. Given the state of Y, we wish to obtain the state of X. Furthermore, the two fields X and Y are coupled through a dependency (called *emission distribution*) where the parameters of the distribution of the random variable y_i depend on the label assigned to the random variable x_i. It may be recalled that the relation between y_i and x_i is

$$f(y_i|x_i = \ell) = \frac{1}{\sqrt{2\pi\sigma_\ell^2}}\ exp\ \{-\frac{(y-\mu_\ell)^2}{2\sigma_\ell^2}\} \tag{3.8}$$

In general, the image model must capture the physical reality as much analytically and computationally as possible. The physical reality that is captured by the image i.e. the particle projections exhibit a spatial coherence. In order to address this aspect of the physical reality, the field X is assumed to be an MRF. The image model must also capture adequately the noise that is introduced in the image capture process. The model captures this by way of assuming the emmission distribution as a normal distibution.

It is also assumed that given the label x_i of X_i, the value taken up by the random variable Y_i is independent of the values of other X_i's

$$P(y|x) = \prod_{i \in S} P(y_i|x_i) \tag{3.9}$$

Hence

$$p(y_i|x_{N_i}, \theta) = \sum_{\ell \in L} f(y_i|x_i = \theta_\ell)\ p(\ell|X_{N_i}), \tag{3.10}$$

where $\theta = \{\theta_\ell\ =\ \{\mu_\ell, \sigma_\ell\},\ \forall \ell \in L\}$.

The Figure 3.4 illustrates the relationship between the label field X and the image field Y. A further illustration of the image model is shown on a identified

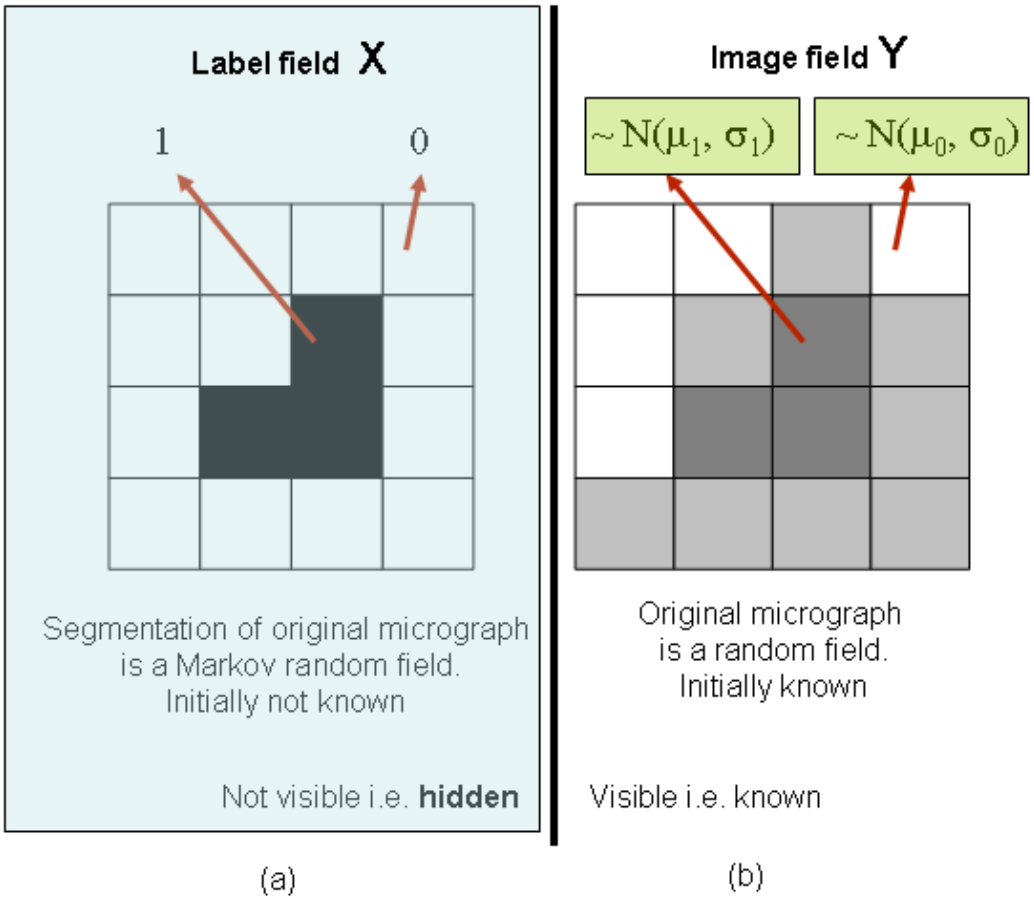
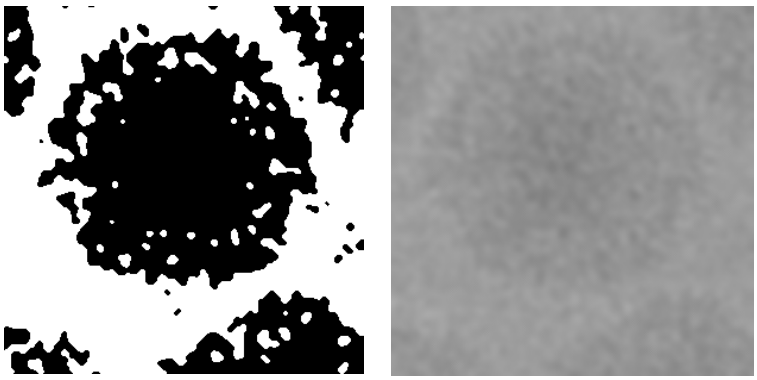


Figure 3.4: An illustration of the two fields involved in the image model. The field on the left corresponds to the label field – X i.e. the segmented image and it's realization is not known i.e. hidden initially. The field on the right corresponds to the image field Y – i.e. the original micrograph image. This field is visible. The shaded areas are not known initially i.e. the field X and the parameters $\mu_0, \sigma_0, \mu_1, \sigma_1$ of the emission distribution.



a b

Figure 3.5: A further illustration of the relation between the image field and its corresponding label field as shown for a identified particle projections a) The label field – X (hidden). b) The image field – Y (visible)

particle projection in the Figure 3.5 (a) and 3.5 (b). Labels from the set {0,1} have been assigned to the pixels of X in Figure 3.4(a). For any pixel X_i in the label field X the intensity at the corresponding pixel Y_i in Y follows a Gaussian distribution with its parameters indexed by the label. i.e. X_i "emits" Y_i. Thus, the intensities in the image field Y are distributed according to a Gaussian distribution with parameters from the set $\{(\mu_0, \sigma_0),\ (\mu_1, \sigma_1)\}$ depending whether X_i is 0 or 1.

3.1.4 MRF Estimation

The aim is to obtain the most likely realization of the segmentation field X given the observed field Y (the micrograph). If we represent the true labelling of the MRF X by $\hat{x}$ and an estimate of it by x^*, then the *Maximum A Posteriori* (MAP) estimate of x can be given by

$$\hat{x} = \underset{x \in X}{arg\ max}\ \{P(y|x)P(x)\} \tag{3.11}$$

But we know that

$$P(x) = \frac{1}{Z}\ exp(-U(x)) \tag{3.12}$$

and

$$P(y|x) = \prod_{i \in S} p(x_i|y_i) = \frac{1}{\sqrt{2\pi}}\ exp\ (-\frac{(y_i - \mu_\ell)^2}{2\sigma_\ell^2} - log(\sigma_{x_i})) \tag{3.13}$$

Hence maximizing $P(y|x)P(x)$ is equivalent to minimizing

$$U(x) + U(y|x) \tag{3.14}$$

where,

$$U(y|x) = \sum_{i \in S} [\frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2} + log(\sigma_{x_i})] \tag{3.15}$$

Hence we need to find particular values for the field of random variables X_i such that the above function is minimized. A semi optimal solution for minimization of this expression is obtained using the *iterated conditional modes*(henceforth I.C.M.) algorithm proposed by Besag [6]. The algorithm is based upon an iterative local minimization strategy where given the data, y, and the other labels $x_{S-i}^{(k)}$, the algorithm sequentially updates each $x_i^{(k)}$ into $x_i^{(k+1)}$ by minimizing $U(x_i|y, x_{S-i})$.

3.1.5 ICM

The ICM is an algorithm based on an iterative local maximization strategy [37]. For the data d and the other labels $f_{S-i}^{(k)}$, the algorithm sequentially updates each $f_i^{(k)}$ into $f_i^{(k+1)}$ by maximizing $P(f_i|d, f_{S-\{i\}})$, with respect to f_i. The assumptions made in calculating $P(f_i|d, f_{S-\{i\}})$ are

1. The observation components $d_1,\ d_2,\ \ldots,\ d_m$ are conditionally independent given f and each d_i has the same known conditional density function $p(d_i|f_i)$ dependent only on f_i.

2. Markovianity of labels.

It follows from these assumptions that

$$P(f_i|d, f_{S-\{i\}}) \propto p(d_i|f_i)\ P(f_i|f_{N_i}) \tag{3.16}$$

Maximizing equation 3.16 is equivalent to minimizing the corresponding posterior potential using the following rule

$$f_i^{(k+1)} \leftarrow arg\ \underset{f_i}{min} V(f_i|d_i, f_{N_i}^{(k)}) \tag{3.17}$$

where

$$V(f_i|d_i, f_{(N_i)}^{(k)}) = \sum_{i' \in N_i} V(f_i|f_{i'}^{(k)}) + V(d_i|f_i) \tag{3.18}$$

The iteration continues until convergence. The convergence is guaranteed for the serial updating [6]. The result obtained by ICM depends very much on the initial estimator.

3.2 Expectation Maximization

Our knowledge being incomplete as regards the realization of the label field X and the parameters of the emission distribution μ_0, σ_0, μ_1 and σ_1. Expectation maximization [13] (henceforth EM) is a standard technique to estimate the parameters of a model when the data available are insufficient or incomplete. The data is clearly incomplete in image model as we do not know the realization of X and the parameters of the emission distribution. We start with an initial guess-estimate of the parameters and obtain an estimate of the "incomplete data" using these parameters. Once the "complete data" (along with some artificial data points) become available, the parameters of the model are estimated again to maximize the probability of occurrence of the "complete data". Successive iterations result in increasingly more refined estimates of the parameters.

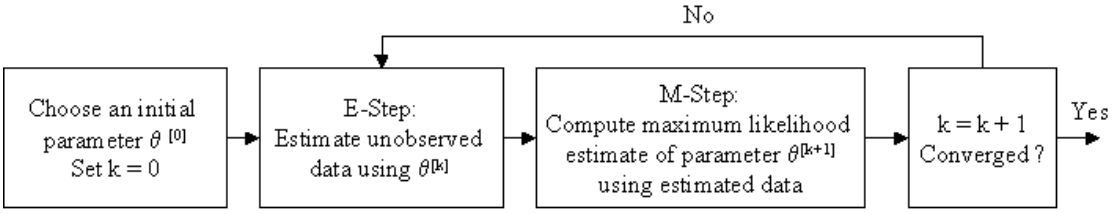


Figure 3.6: A schematic for expectation maximization algorithm.

Let $\theta^{(0)}$ be the initial estimate of the parameters of the model. The E.M. algorithm consists of two major steps [47]. Let Y denotes the sample space of the observations, and let $\mathbf{y} \in R^m$ denote an observation from Y. Let Γ denote the underlying space and let $\mathbf{x} \in R^n$ be an outcome from Γ with $m < n$. The data $\mathbf{x}$ is referred to as the *complete data*. The complete data $\mathbf{x}$ is not observed directly, but only by means of $\mathbf{y}$, where $\mathbf{y} = \mathbf{y}(\mathbf{x})$, and $\mathbf{y}(\mathbf{x})$ is many-to-one mapping. An observation $\mathbf{y}$ determines a subset of Γ, which is denoted as $\chi(\mathbf{y})$. The probability

density function (pdf) of the complete data is $f_X(\mathbf{x}|\theta)$, where $\theta \in \Theta \subset R^r$ is the set of parameters of the density. The pdf f is assumed to be a continuous function of θ and appropriately differentiable. The Maximum Likelihood estimate of θ is assumed to lie within the region Θ. The pdf of incomplete data is

$$g(\mathbf{y}|\theta) = \int_{\Gamma(y)} f(\mathbf{x}|\theta) d\mathbf{x} \tag{3.19}$$

let $l_y(\theta) = g(\mathbf{y}|\theta)$ denote the likelihood function and let $L_y(\Theta) = log\ g(\mathbf{y}|\theta)$ denote the log-likelihood function. The basic idea behind the EM algorithm is that we should like to find θ to maximize $log\ f(\mathbf{x}|\theta)$, but we do not have the data $\mathbf{x}$ to compute the log-likelihood. So instead we maximize the expectation of $log\ f(\mathbf{x}|\theta)$ given the data $\mathbf{y}$ and our current estimate of θ. This can be expressed in two steps. Let $\theta^{[k]}$ be our estimate of the parameters at the kth iteration. For the E-step compute :

$$Q(\theta|\theta^{[k]}) = E[log\ f(\mathbf{x}|\theta)|\mathbf{y}, \theta^{[k]}] \tag{3.20}$$

It is important to distinguish between the first and the second arguments of the Q functions. The second argument is a conditioning argument to the expectation and is regarded as fixed and known at every E-step. The first argument conditions the likelihood of the complete data. For the M-step let $\theta^{[k-1]}$ be that value of θ which maximizes $Q(\theta|\theta^{[k]})$: $\theta^{[k+1]} = \underset{\theta}{arg\ max}\ \ Q(\theta|\theta^{[k]})$. It is important to note that the maximization is with respect to the first argument of the Q function, the conditioner of the complete data likelihood. The EM algorithm consist of choosing an initial $\theta^{[k]}$, then performing the E-step and the M-step successively until convergence. Convergence may be determined by examining when the parameters quit changing, i.e. stop when $\|\theta^{[k]} - \theta^{[k-1]}\| < \epsilon$ for some ϵ and some appropriate norm $\|.\|$.

In summary,

Step 1 - Expectation – Calculate the expectation with respect to the unknown underlying parameters using the current estimate of the parameters, conditioned on the observations

$$Q(\theta|\theta^{[k]}) = E[log\ f(\mathbf{x},\mathbf{y}|\theta)|\mathbf{y},\theta^{[k]}] = \sum_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}|\mathbf{y},\theta^{[k]})\ log\ f(\mathbf{x},\mathbf{y}|\theta) \quad (3.21)$$

Where, k is an variable representing a particular iteration of the algorithm, Q is a function of the parameters set θ conditioned on the parameter set obtained in the previous iteration and $E(\cdot)$ is the expectation function.

Step 2 - Maximization – Calculate the new estimate of the parameters

$$\theta^{[k+1]} = arg\ \underset{\theta}{max}\ Q(\theta|\theta^{[k]}) \quad (3.22)$$

3.2.1 EM for image model

The equations describing the application of E.M. to the image model are

$$\mu_\ell^{(k+1)} = \frac{\sum_{i\in S} P^{(k)}(\ell|y_i)y_i}{\sum_{i\in S} P^{(k)}(\ell|y_i)} \quad (3.23a)$$

and

$$(\sigma_\ell^{(k+1)})^2 = \frac{\sum_{i\in S} P^{(k)}(\ell|y_i)(y_i-\mu_\ell)^2}{\sum_{i\in S} P^{(k)}(\ell|y_i)} \quad (3.23b)$$

where

$$P^{(k)}(\ell|y_i) = \frac{g^{(k)}(y_i;\theta_\ell)P^{(k)}(\ell|x_{N_i})}{p(y_i)} \quad (3.23c)$$

$P^{(k)}(\ell|x_{N_i})$ involves the MAP estimation as described earlier. The intermediate steps are described in detail in [87].

The refinement may be seen as an iterative optimization procedure where the parameters, namely μ_0, σ_0, μ_1, σ_1, and labels of the label field X are estimated using the E.M. algorithm. At this point it must be reemphasized that the iterative step of E.M. includes I.C.M. [6] which, as described in section 3.1.5, is a local optimization algorithm.

CHAPTER 4

HMRF BASED SEGMENTATION ALGORITHM AND PARTICLE BOXING

Be not ashamed of mistakes and thus make them crimes.

— *Confucius.*

Segmentation is a low level operation that is principally based on the pixel level intensities. The aim of segmentation is to delineate those parts of an image that correspond to distinct objects in the real world that are depicted in the image. Intuitively this would require the use of any knowledge one has about certain properties of the representation of the objects in the image that is being segmented. However, sometimes certain types of images can be segmented successfully without much knowledge about the properties of the object's representation in the image. Segmentation must be decoupled from the higher level particle recognition step i.e. particle boxing so that different particle boxing methods can be used with various segmentation methods. This decoupling gives us the flexibility of changing the particle boxing method as and when better algorithms become available. However, segmentation methods that use the features of the object that is being recognized achieve a better segmentation.

We begin by describing the preprocessing of micrographs using anisotropic diffusion. This is followed by a description of a Markov random field based image

segmentation algorithm. The particle boxing step is described in the concluding section.

4.1 Preprocessing and segmentation

One of the trade-offs in imaging biological particles using cryo TEM is that of using a high dosage of electrons vis-a-vis keeping the structure of the macromolecule intact. At high dosage levels ($>$ about 12 $e^-/\text{Å}^2$) the particles incur damage and the images thus obtained are of not much use. Hence, in order to minimize the damage during the imaging process, micrographs are imaged at a very low dosage ($<$ about 10 $e^-/\text{Å}^2$). The use of such low dosage leads to a very low signal to noise ratio (SNR) of the micrograph images. The SNR values of micrographs typically vary from 0.3 to 0.08. Such low SNR values make the problem of identification of the projections quite difficult. The segmentation step is severely affected by such low SNR values. In order to achieve an improvement in the SNR various filtering techniques, both during imaging and during the image processing phase are commonly used. Here we describe the filtering, as a preprocessing step, that is performed in the image processing phase.

4.1.1 Preprocessing

The preprocessing of micrographs is done so that the particle identification algorithm that follows it would be able to perform better. The various filtering techniques use low-pass filtering to allow only certain frequency to be passed to the output. It is fairly assumed that the noise exists mainly at the higher fre-

quencies. Techniques such as gaussian filtering and anisotropic diffusion etc. are common. However, preprocessing comes at the cost of altering the signal since that part of the signal that exists at higher frequencies is not present in the filtered output due to attenuation of higher frequencies. The anisotropic diffusion also shows such a behavior due to it's equivalence to gaussian filtering which essentially is a low pass filter.

4.1.2 Anisotropic Diffusion

Often, an automatic particle identification method involves a pre-processing step designed to improve the signal to noise ratio of a noisy micrograph. Various techniques, such as histogram equalization, and different filtering methods are commonly used. Now we describe briefly an anisotropic filtering technique we found very useful for enhancing the micrographs before the segmentation and labelling steps.

While other pre-processing techniques such as histogram equalization, attempt to increase the dynamic range of the low-contrast micrographs, the anisotropic diffusion in fact reduces this dynamic range, as seen in Figure 4.1 (c). The aim of anisotropic diffusion is to enhance the "edges" present in the image by smoothing the regions devoid of 'edges'.

A diffusion algorithm is used to modify iteratively the micrograph, as prescribed by a partial differential equations (PDE) [57]. Consider for example the isotropic diffusion equation

$$\frac{\partial I(x,y,t)}{\partial t} = div(\nabla I) \tag{4.1}$$

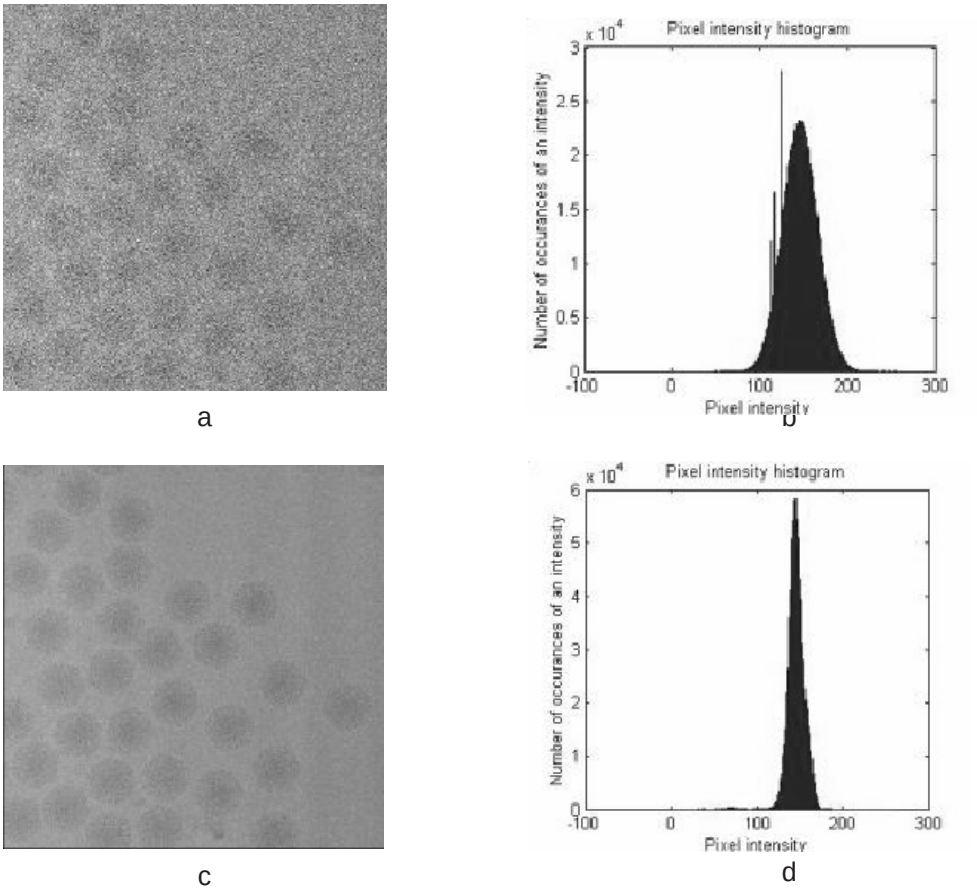


Figure 4.1: Anisotropic filtering. a) A portion of a micrograph of frozen-hydrated Ross River virus particles. b) Histogram of the pixel intensities in the micrograph displayed in (a). c) The image in (a) after 10 cycles of anisotropic filtering. d) Histogram of the pixel intensities in the micrograph displayed in (c).

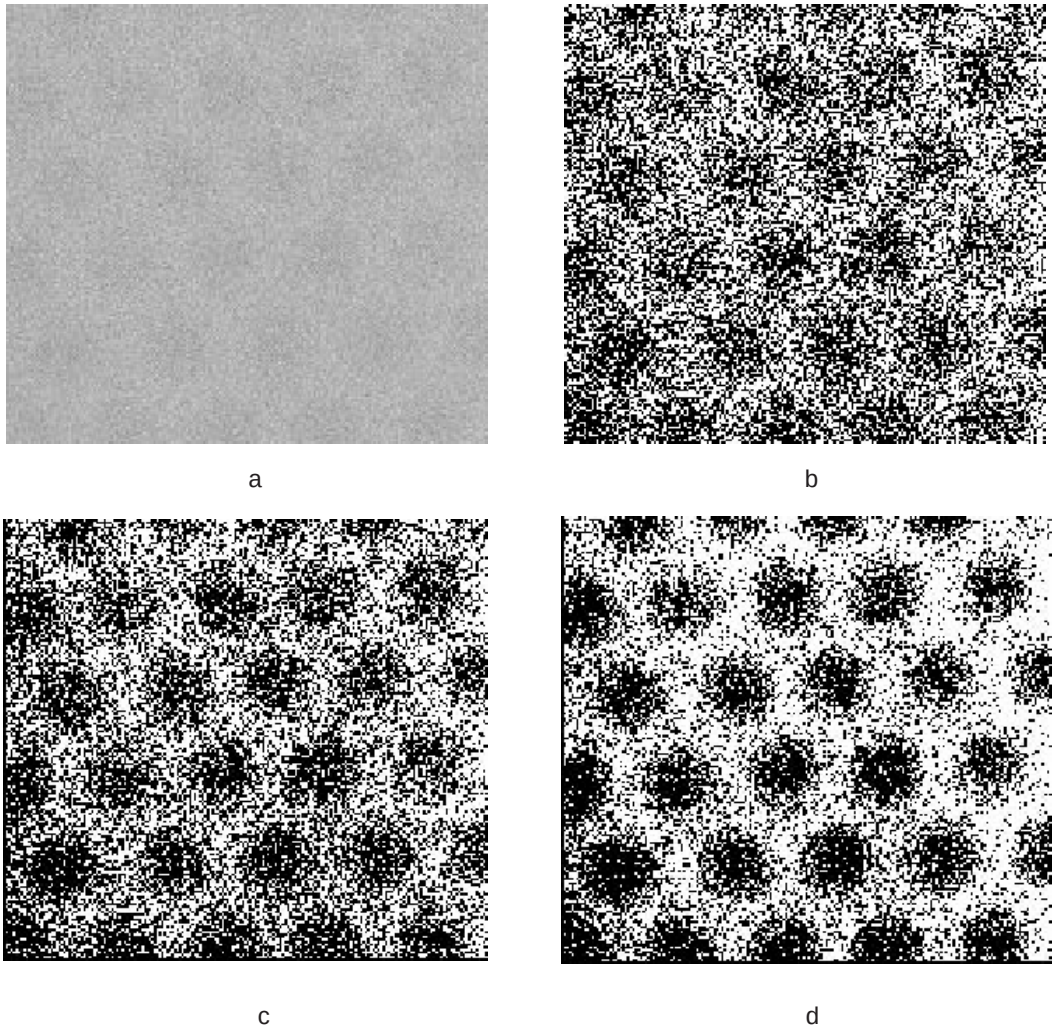


Figure 4.2: The effect of the number of iterations of anisotropic diffusion on segmentation, when all other parameters are kept constant. a) A portion of a micrograph of frozen-hydrated Chilo Iridescent virus (CIV) particles. b) Segmentation without anisotropic diffusion, c) 4 iterations of anisotropic diffusion followed by segmentation d) 10 iterations of anisotropic diffusion followed by segmentation

In this partial differential equation t specifies an artificial time and ∇I is the gradient of the image. Let $I(x, y)$ represent the original image. We solve the PDE with the initial condition $I(x, y, 0) = I(x, y)$, over a range of $t \in \{0, T\}$. As a result, we obtain a sequence of Gaussian filtered images indexed by t.

Unfortunately, this type of filtering produces an undesirable blurring of the edges of objects in the image. Perona and Malik [57] replaced this classic isotropic diffusion equation with

$$\frac{\partial I(x, y, t)}{\partial t} = div[g(\|\nabla I\|)\nabla I] \tag{4.2}$$

where $\|\nabla I\|$ is the modulus of the gradient and $g(\|\nabla I\|)$ is the "edge stopping" function chosen to satisfy the condition $g \to 0$ when $\|\nabla I\| \to \infty$.

The modified diffusion equation prevents the diffusion process across edges. As the gradient at some point increases sharply, signalling an edge, the value of the "edge stopping" function becomes very small making $\frac{\partial I(x,y,t)}{\partial t}$ effectively zero. As a result, the intensity at that point on the edge of an object is unaltered as t increases. This procedure ensures that the edges do not get blurred in the process.

The result of applying ten iterations of anisotropic filtering to an electron micrograph of frozen-hydrated dengue virus particles is illustrated in Figure 4.1. The efficacy of anisotropic diffusion is ascertained by the illustration in Figure 4.2. Clearly a higher number of iterations benefits segmentation. However increasing the number of iterations is detrimental as it causes widespread diffusion resulting in joining of nearby projections in the segmentation output.

4.1.3 Segmentation

Following is a stepwise description of the micrograph segmentation algorithm [72].

1. The image is split into rectangular blocks that are roughly twice the size of the projection of a particle. This is done to reduce the gradient of the background across each processed block. A high gradient degrades the quality of the segmentation process carried out in Step 3. The gradient of the background affects the algorithm in the following ways

   (a) The initialization is based solely on intensity histogram which encode only the frequencies of occurrence of any pixel intensity in the image. Due to the presence of a gradient, the contribution to the count of an intensity for example, may come from the background of a darker region, as well as from the inside of a projection of a virus in a brighter region. When the initialization is done for the entire image it performs poorly, as seen in Figure 4.3.

   (b) The parameters μ_0, σ_0, μ_1, σ_1 are fixed for an image. However, as illustrated in Figure 6.1, they are not the true parameters for intensity distribution across the whole image. The means and variances of pixel intensities are significantly different across the image due to the presence of the gradient. Cutting the image into blocks ensures a lack of drift in these parameters across each block.

2. As a pre-processing step individual blocks are filtered by means of anisotropic diffusion. Such filtering ensures that "edges" are preserved and less affected by smoothing (see Figure 4.1(a) and 4.1(b)). The edge stopping function is

$$g(\nabla I) = e^{-(\|\nabla I\|/K)^2} \tag{4.3}$$

For each block we run a few iteration of the algorithm. $K = 0.2$ to ensure the stability of the diffusion process.

3. The blocks filtered through the anisotropic diffusion based filter are segmented using the HMRF method. The following steps are taken for segmentation

   (a) The initialization of the model is done using a discriminant measure based thresholding method proposed in [55]. The threshold is found by minimizing the intra-class variances and maximizing the inter-class variances. The resulting threshold is optimal [55]. Pixels with intensities below the threshold are marked with the label 1 indicating that they belong to the particle projection. The remaining pixels are marked with the label 0. This initialization is refined using the MRF based model of the image in Step 3(b).

   (b) To refine the label estimates for each pixel within the MAP framework, we use the expectation maximization algorithm. A second order neighborhood, as the one in Figure 3.3, is used. To compute the potential energy functions for cliques we use a Multi Level Logistic (MLL) model. Four iterations of the algorithm are run for each block. The result of the segmentation is a binary image with one intensity for the particle projection and the other for the background.

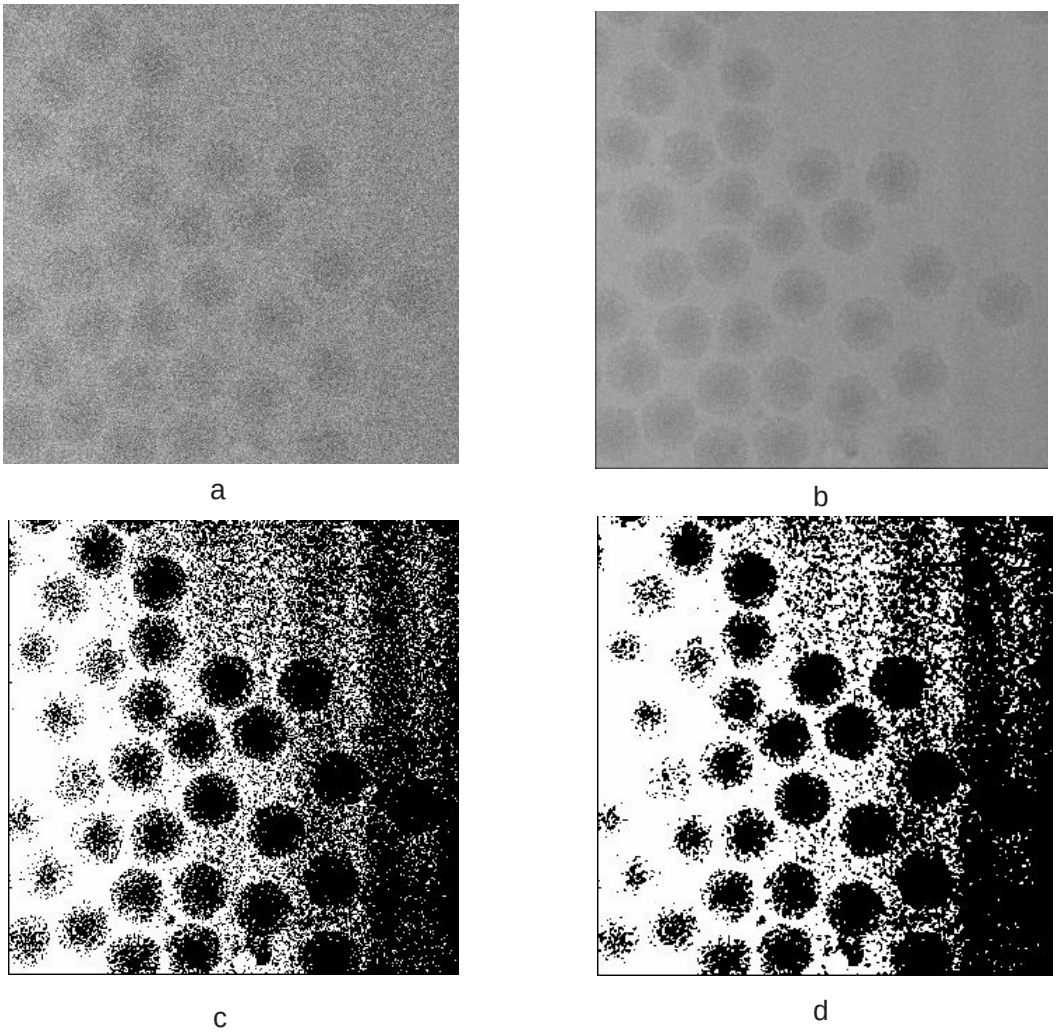


Figure 4.3: The effect of initialization for a micrograph with a pronounced gradient. a) A portion of an original micrograph of frozen-hydrated Ross River virus. b) The micrograph after anisotropic diffusion. c) The effect of the initialization for the HMRF segmentation algorithm for the micrograph in (a). d) The segmented micrograph.

4.2 Boxing Strategies

The goal of particle identification from segmented images of micrographs is to picking regions that correspond to the projections. Boxing means to construct a rectangle with a center co-located with the center of the particle. Once the segmented image is obtained, with or without the use of a low resolution 3D structure of the macromolecule, any of the several boxing strategies detailed below can be used to isolate the regions on the micrograph that correspond to a particle projection. Boxing represents a crucial step in the particle identification process. The decoupling of segmentation and boxing provides us with the advantage of making any of the different combinations based on the structural properties under consideration and quality of the micrograph.

Boxing particles with unknown symmetry is considerably more difficult than the corresponding procedure for iscosahedral particles. First, the center of a projection is well defined in case of icosahedral particle, while the center of the projection of an arbitrary 3D shape is more difficult to define. Second, the result of pixel labeling, or segmentation, is a shape with a vague resemblance to the actual shape of the projection. Typically, it consists of one or more clusters of marked pixels, often disconnected from each other, as we can see in Figure 3.5.

The post-processing of the segmented image to achieve boxing involves morphological filtering operations of opening and closing [74]. These two morphological filtering operations are based on the two fundamental operations called dilation and erosion. For a binary labeled image, dilation is informally described as taking each pixel with value 1 and setting all pixels with value 0 in its neighborhood to the value 1. Correspondingly, erosion means to take each pixel with value 1 in the neighborhood of a pixel with value 0 and re-setting the pixel value

to 0. The term "neighborhood" here bears no relation to the "neighborhood" in the framework of Markov Random Field described earlier in section 3.1. Pixels marked say as "1," separated by pixels marked as "0" could be considered as belonging to the same neighborhood if they dominate a region of the image. The opening and closing operations can then be described as erosion followed by dilation and dilation followed by erosion, respectively.

The decision of whether a cluster in the segmented image is due to noise, or due to the projection of a particle is made according to the size of the cluster. For an icosahedral particle, additional filtering may be performed when the size of the particle is known. Such filtering is not possible for particles of arbitrary shape. A fully automatic boxing procedure is likely to report a fair number of false hits, especially in very noisy micrographs.

4.3 Results

We present four images: the original, the image after preprocessing, the segmented image, and the original image with the particles boxed.

Table 4.1: The quality of the solution provided by the HMRF particle identification algorithm for a micrograph of the Chilo Iridescent virus (CIV)

Number of particles detected manually	Number of particles detected by MRF	False Positives	False Negatives
277	306	55	26

Automatic identification of particle projections from a micrograph requires that the results be obtained reasonably fast. Hence, in addition to analysis pertinent to the quality of the solution, we report the time required by the algorithm

for different size and number of particles in a micrograph. Table 4.2 lists the time devoted to different phases of our algorithm and demonstrates that pre-processing and segmentation account for 9799 % of the computing time.

Table 4.2: Time in seconds for the main processing steps of the of HMRF particle identification algorithm

Image Size	Anisotropic filtering	MRF Segmentation	Post-processing	Total
1174 x 940	17	23	2	42
5457 x 6000	560	744	28	1332
8768 x 11381	2102	3012	142	5256

4.4 Parallelization

The aim of reaching atomic resolution limits with cyro-EM based single particle analysis leads to the need for automatic methods for identification of hundreds of thousands of particle projections. In general, the processing of micrographs for particle identification for such huge numbers of projections is a highly computation intensive task. The performance data for the sequential algorithm reported in section 4.3 indicate that the execution time of the segmentation process is prohibitively large. Considering the fact that in future there may be a need to process even larger micrographs and that we may need to improve the quality of the solution it becomes abundantly clear that we should consider parallel algorithms for automatic particle identification ([42] and [73]).

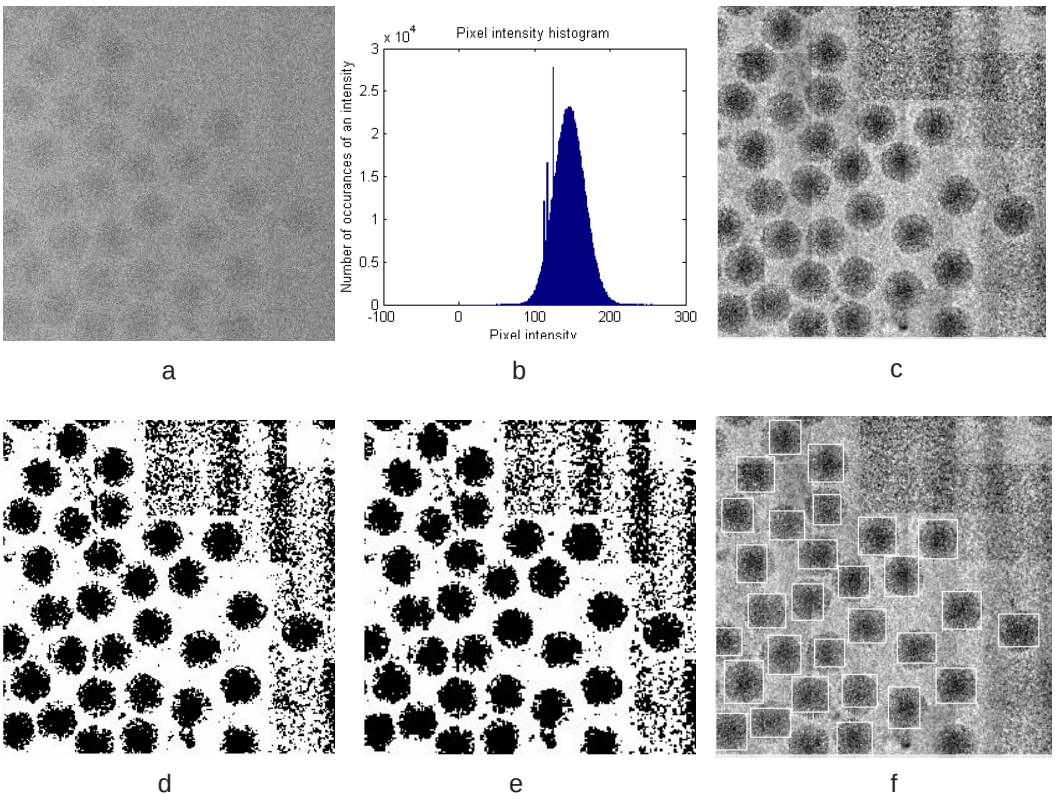


Figure 4.4: a) Portion of a micrograph of frozen-hydrated Ross River virus. b) The histogram of the pixel intensities for the image in (a). c) The micrograph after anisotropic diffusion filtering. d) The micrograph after the initialization step of HMRF. e) Segmented micrograph. f) Boxed particles in the micrograph.

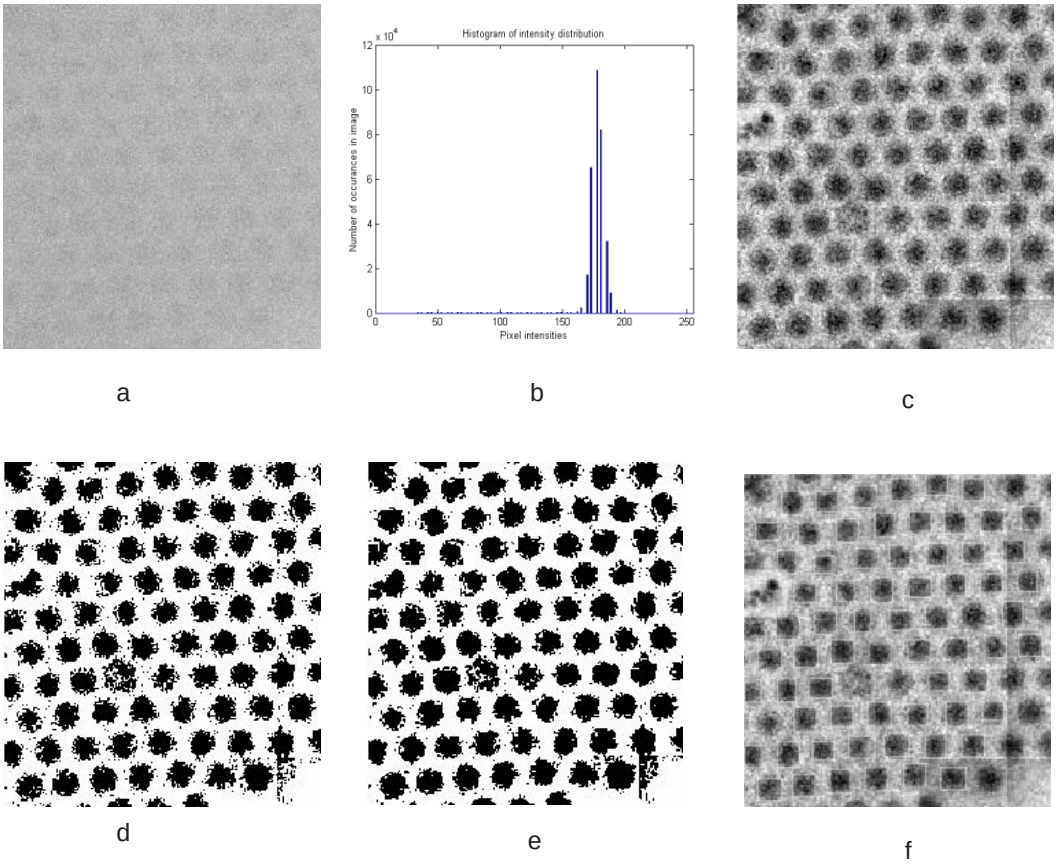


Figure 4.5: a) A portion of a micrograph of frozen hydrated sample of Chilio Iridescent virus (CIV). b) The histogram of the pixel intensities for the image in (a). c) The micrograph after anisotropic diffusion filtering. d) Micrograph after the initialization step of HMRF. e) Segmented micrograph. f) Boxed particles in the micrograph.

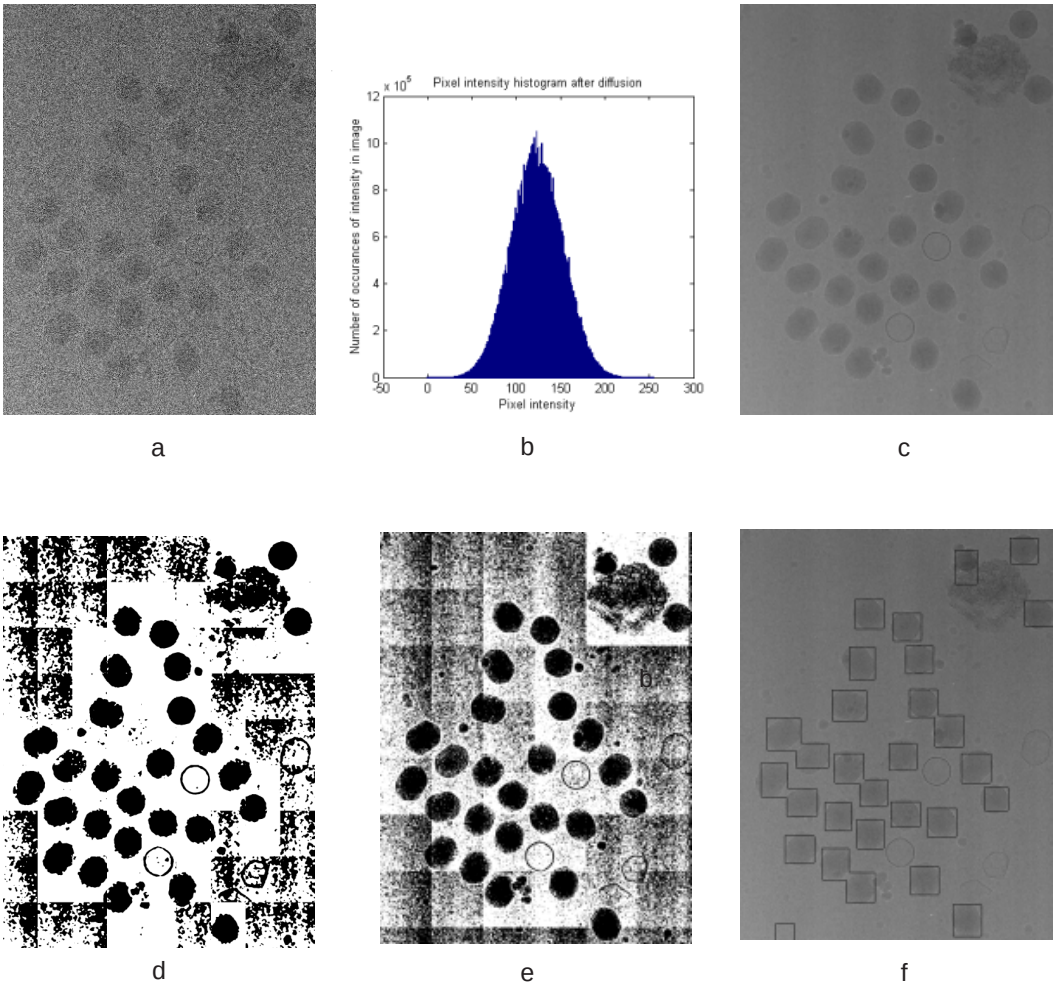


Figure 4.6: a) Portion of a micrograph of frozen-hydrated bacteriophage T4 prolate virus. The virus does not have icosahedral symmetry. b) The histogram of the pixel intensities for the image in (a). c) The micrograph after anisotropic diffusion filtering. d) Micrograph after the initialization step of HMRF. e) Segmented micrograph. f) Boxed particles in the micrograph.

4.4.1 Background

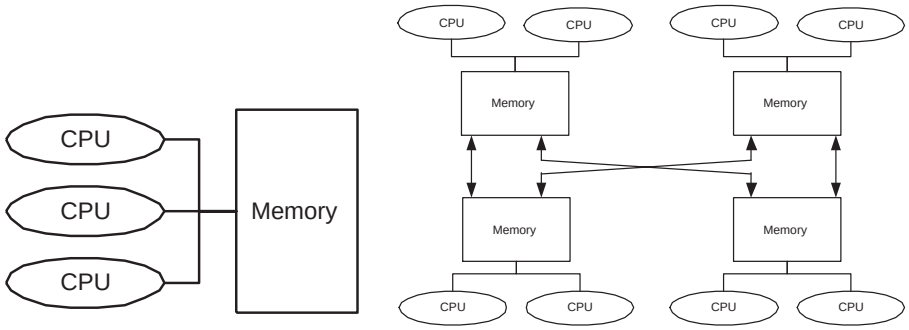
Parallel architectures have traditionally been classified using the Flynn taxonomy [16] as SISD, SIMD, MISD, and MIMD. The MIMD category is very broad and is typically further classified into categories based on memory organization.

4.4.2 Shared memory architecture

All processes share the same address space. The inter process communication takes place via shared variables. There are two major classes of shared memory systems.

- SMP (Symmetric multiprocessor) - All processors share a connection to the common memory. The memory access speeds are equal for all the processors. SMP systems do not scale well.
- NUMA (Non Uniform Memory Access) - NUMA systems exhibit a non uniform access to the memory from the processors. Some blocks of memory may be closer to some processors than others.

The SMP architecture affords an ease of programming for the programmer when compared to the NUMA architecture. With the NUMA architecture, the programmer needs to care about the locality of the data. In order to overcome the non locality (and hence the non uniformity) of memory access, each processor has a cache and an associated cache coherence protocol.



(a) (b)

Figure 4.7: Classification of MIMD parallel systems based on memory access. The Symmetric Multiprocessor (SMP) architecture is depicted in (a) and the non uniform memory access (NUMA) architecture is depicted in (b).

4.4.2.1 Distributed memory architecture

Each process has its own address space and the communication between processes takes place through message passing.

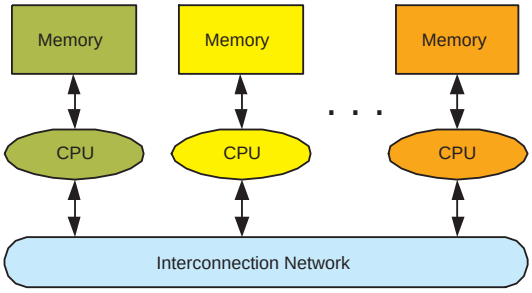


Figure 4.8: Classification of MIMD parallel systems based on memory access. The Symmetric Multiprocessor (SMP) architecture is depicted in (a) and the non uniform memory access (NUMA) architecture is depicted in (b).

4.4.3 Design

The design of a parallel algorithm involves various factors regarding both the problem at hand and the system for parallelization. Some of the things that must be considered when designing a parallel algorithm are the following -

1. Dependence of operations for the given algorithm.
2. Architecture of the parallel system.

The problem of segmentation by the algorithm mentioned above is embarrassingly parallel because there is almost no dependence of any computational step. Hence the design of a parallel algorithm is fairly straightforward. A scheme for parallelization of the algorithm is given in Figure 4.9.

4.4.4 Parallel Algorithm

For an algorithm designed for a cluster of PCs, we divide the image into blocks, where the size of the blocks is roughly twice the estimated size of the particles. The size of the particle is a user input parameter which needs to be provided at only once. Let the image be divided into m by n blocks. Further, let there be p processors. A pseudo code for the parallel algorithm for segmentation is shown below.

```
if master node
 then
   divide the image into m by n blocks
   for i = 0 to m-1
```

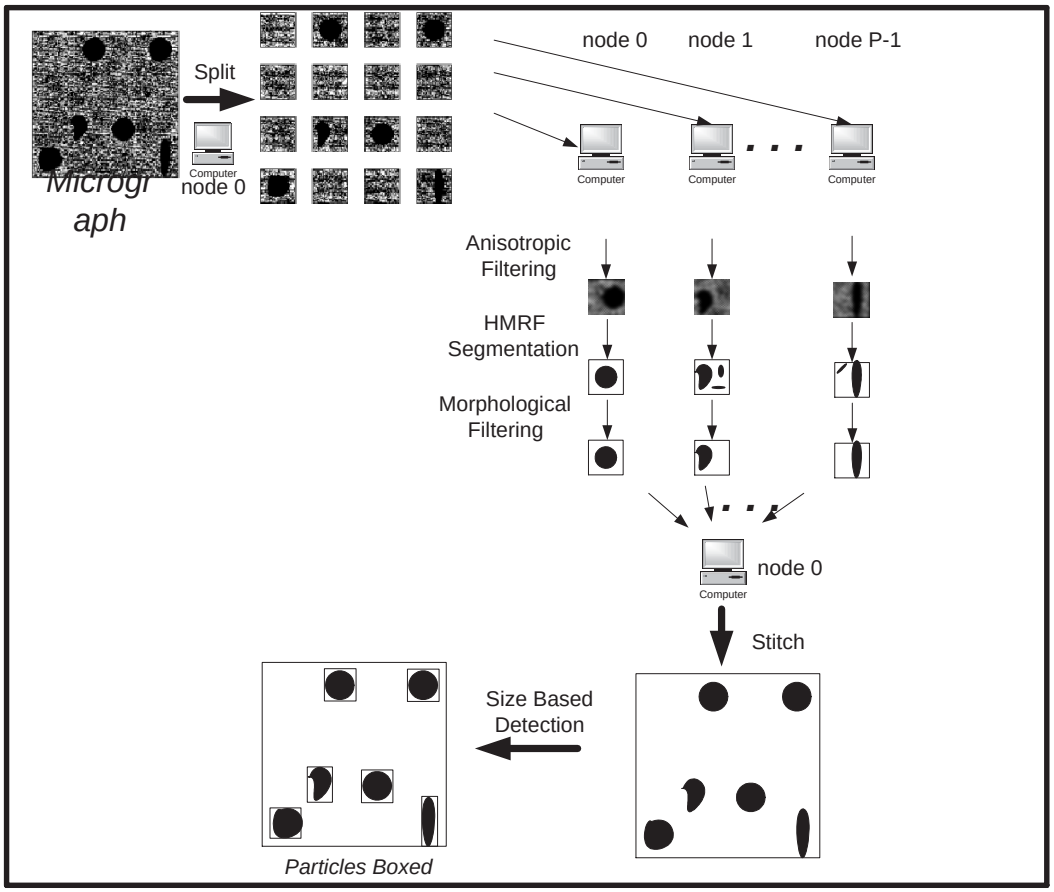


Figure 4.9: Schematic of the parallel version of HMRF Segmentation algorithm.

```
     for j = 0 to n-1
       send block(i,j) to node (i+j) mod p
     end for
   end for
receive the block/s from the master node

filter the block/s received using anisotropic diffusion

segment the block/s received

perform morphological filtering on the block/s

send the block/s to the master node

if  master node
 then
   while not all blocks are received
     receive the next block
   end while
   Stitch all the blocks into final segmented image
```

The computation was performed on a cluster of 44 dual processing (2.4 GHz Intel™Xeon processors) nodes, each with 3 GB of main memory and interconnected using 1 Gbps ethernet. The processing time significantly reduces with an increase in processing nodes until about 20 nodes, after which the processing time does not decrease significantly with an increase in the number of processing nodes. The result of execution of the above algorithm on three different sizes of

images viz. (in pixels) 680x750, 1356x1492 and 5424x5968 is shown in Figure 4.4.4. As can be seen from the figure, the speedup starts to level off with an increase in the number of nodes reflecting the fact that communication costs start playing a more significant role as the number of nodes is increased. However this levelling off is not observed for the image of size 5424x5968 since communication time becomes less significant as the image size, and hence the computation time is increased. The vertical bars in Figures 4.10(b), 4.10(c) and 4.10(d) correspond to the 95% confidence interval. The image of size 5424x5968 corresponds to the image shown in Figure 4.5. The confidence intervals for the image of size 680x750 are very narrow indicating a high accuracy of the measurement. The experiments were conducted on a cluster simultaneously with other unrelated processes running. We believe that the difference in size of confidence intervals for the mean of the measured execution time for the three image sizes may be due to this fact.

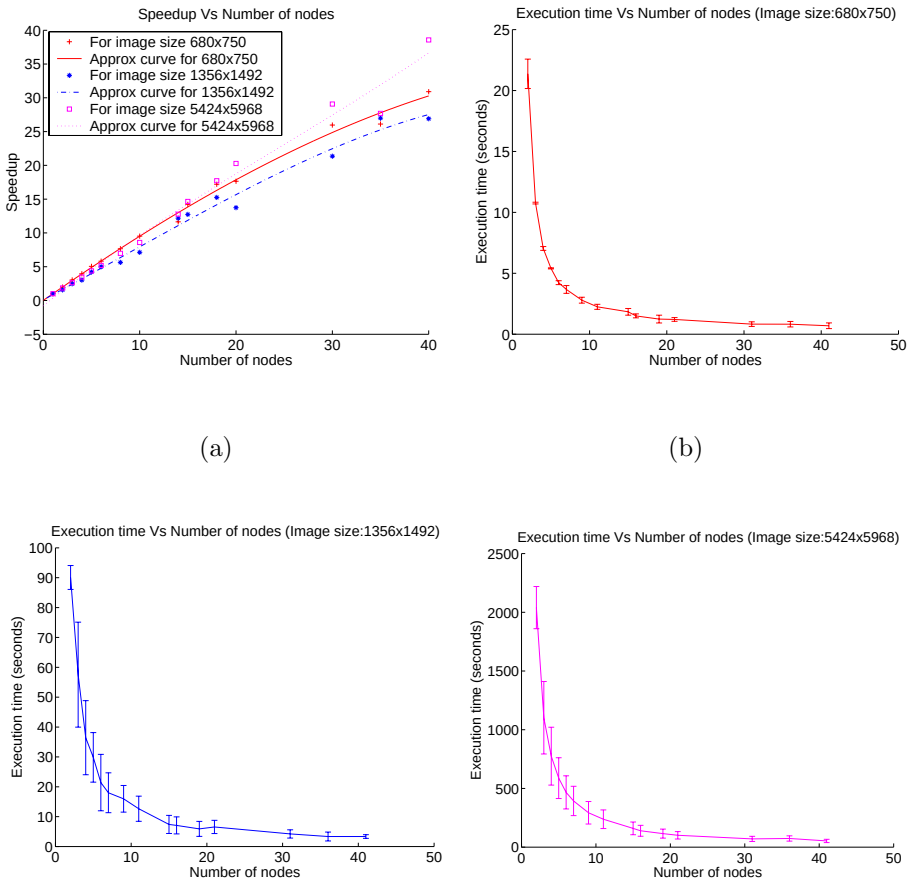


Figure 4.10: A plot of speedup vs Number of nodes for three different sizes of micrograph images is shown in (a). Plots of *Total Computation time* Vs *Number of nodes* for three different sizes of micrograph images (in pixels) 680×750, 1356×1492 and 5424×5968 are shown in (b), (c) and (d) respectively. The computation was performed on 44 dual processing nodes (2.4 GHz Intel™Xeon) each with 3Gb of main memory and connected using 1 Gbps. The vertical bars indicate a 95% confidence interval, computed using 30 readings corresponding to each value of *Number of nodes*.

CHAPTER 5

VIRTUAL WORKBENCH

Everything should be made as simple as possible, but no simpler.

— *Albert Einstein.*

5.1 Motivation

Automatic identification of particle projections in micrographs is critical for high-resolution reconstruction in Transmission Electron Microscopy (TEM). Manual identification of tens of thousands of particle projections is impractical, it is time-consuming and prone to errors. A high-resolution reconstruction requires not only a large number of particle projections, but also high-quality input data; we have to minimize the number of false hits. A fair number of algorithms and methods for automatic particle identification have been proposed. However, the performance of these algorithms varies with various factors such as the contrast of the image, the gradient of the background, the shape of the macromolecules, and other factors related to the experimental setup used for data collection. Currently it is difficult to compare the quality of the solution produced by various algorithms for automatic particle identification. We have developed, for the first time, a collaborative tool for the cryo-TEM community, which in short term will

allow the scientists to decide which algorithm performs best on a specific set of micrographs. It is expected that one should be able to select the best algorithm(s) automatically. Whenever the ground truth for a set of micrographs is available, the system will allow one to make objective statements regarding the performance of different algorithms. The performance of a particle identification algorithm has traditionally been benchmarked against the ground truth i.e. manual pickings by a trained human particle picker, although this metric is not universally accepted [25] as the best method of comparison.

The system has been christened 'Virtual workbench' to distinguish it from traditional benchmarks. The traditional benchmarks consist of a closed (fixed) set of algorithms/programs and input data. The goal of benchmarking a new system is to establish if it is 'better' than a set of existing systems; to do so one would measure an objective performance indicator, e.g. the execution time, without being concerned with "fuzzier" attributes such as the quality of the solution. Benchmarking is traditionally done by an organization on behalf the user community only once, on a prototype system. The Virtual workbench is open-ended, new algorithms and new data shall be included as they become available. The system will be available to a large user community. As more complex algorithms whose performance depends upon a set of parameters are developed, the system will allow a scientist to determine the optimal parameters for an algorithm on a particular set of input data. The system will facilitate the development of new algorithms for automatic identification of particles, such as those which exhibit icosahedral symmetry, as well as particle which exhibit other symmetries and allow a scientist to compare objectively various algorithms and programs.

5.2 Functional Features

The Virtual benchmark essentially provides a service to the research community. It enables, among other things such as collaboration, an ability to benchmark a collection of particle identification algorithms over a representative set of micrographs. The service must possess certain functional features in order to be of utility to the targeted audience. Such functional features are described below.

5.2.1 Characterization

In order to provide the facility of benchmarking and to allow comparison among the particle identification algorithms, the algorithms must essentially be characterized by their performance. The performance of particle identification algorithms is heavily dependent on the quality of the data that is input to them. Hence the input data i.e. micrographs must also be characterized in order to obtain a relation between the performance of particle identification algorithms with the 'quality' of the micrographs. Characterization of algorithms and micrographs is discussed in detail in Chapter 6.

5.2.2 Visualization

Visualization is an essential component of the workbench. Numerous parameters affect the performance of the particle identification algorithms. In order to allow comparisons of the algorithms, visualization of the results of the algorithms and micrograph characteristics must be provided.

5.2.3 Collaborative environment

In addition to providing a benchmarking facility, the virtual workbench also provides a collaborative environment to its users. Such an environment consists of virtual groups whose members can exchange data and results between themselves. In order to exchange data, users can simply allow access to their data for the other members. Since the data is maintained in the virtual workbench, a member to member transfer of data is not required. Nevertheless, access to the micrographs and algorithms must be secure in order to protect the user's data and intellectual property rights.

5.2.4 Miscellaneous features

Apart from the features described above, the workbench also provides the users with the ability to generate and share synthetic data. Moreover the algorithm results can be tested over the synthetic data thus generated. Since the particle positions in the synthetically generated micrographs is known beforehand, the task of estimation of the metrics for performance of the algorithms can be automated. This provides an elegant way to both to test the algorithms and to characterize them. Further details are given in section 6.2.2 of Chapter 6. However since the synthetic data is not truly accurate representations of the micrographs obtained by imaging of the macromolecules, the faithfulness of performance of the algorithms on such data vis-a-vis real micrograph data is closely tied to the process of creation of these synthetic micrographs. The process of creating snthetic micrographs is given in section 6.2.2.

5.3 Design

The workbench essentially consists of a collection of

1. Micrographs,
2. Particle identification algorithms provided by the users and the results of these algorithms on a subset of the micrographs in the collection,
3. Ancillary algorithms for characterization, visualization and analysis,
4. Web accessible tools to explore the collections and selectively perform actions on these collections or some subset of it, and
5. Data regarding users and procedures to enable collaboration.

All micrographs and algorithms provided by the users and the results of these algorithms (on a subset of the micrographs present in the workbench) have meta-information associated with them. The table 5.1[1] gives a non-exhaustive list of the meta-information regarding each micrograph stored in the workbench.

Other meta-information such as details about the research group/individual involved in preparing the micrograph is also associated with the micrograph.

A non-exhaustive list of such metrics includes

1. The expected value and the variance of ratio of particle identified by the method versus the number of 'true' particles.
2. The expected value and the variance of the number of 'false positives'.
3. A measure of the quality of the solution, for example the average error in determining the center of each particle.

[1] This is a sample of a larger set of micrograph meta-information in the actual implementation

Table 5.1: For each micrograph this list of meta information is stored in the database.

Micrograph meta-information	
`Size`	Size of the image in pixels
`NumTrue`	True number of particles
`NoiseMean`	Empirical mean of the noise
`NoiseVar`	Empirical variance of the noise
`EVolt`	Electron voltage of microscope
`CoolMed`	Cooling medium (Liquid He or N)
`DensType`	Type of densitometer (line or point)
`DensResol`	Resolution of the densitometer
`CTFAlgo`	Algorithm used to compute the CTF
`EnvlParam`	Envelope function parameter
`EnvlForm`	Functional form of the envelope function
`DefocusMaj`	Defocus major
`DefocusMin`	Defocus minor
`SHTilt`	Specimen holder tilt
`ImgFrmt`	Format of the image
`ImgReader`	File for reading the image
`ImgWriter`	File for writing the image to a file
`MacromolProjShapeID`	ID for associating shapes with micrographs
`isSynthetic`	Is the micrograph real or synthetic
`FileLoc`	Path to the image file
`isDirty`	Processed status of the micrograph
`ParamsComputed`	None/Some/All parameters computed
`MisPrivate`	Privacy of the micrograph

Table 5.2: For each algorithm the given list of meta information is stored in the database.

Algorithm meta-information	
`Name`	Size of the image in pixels
`PID`	ID of the uploading account holder
`AlgoFileLoc`	Location of the file
`LibID`	ID for the libraries used
`Publication`	Citation of any associated publication
`AisPrivate`	Privacy of the algorithm

Table 5.3: For each result obtained by a run of an algorithm using a unique set of parameters, the results are stored in a results table. A short list of fields of this table is given.

Algorithm results meta-information	
`AlgoID`	ID of the algorithm
`microgID`	ID of the micrograph
`AlgoResFileLoc`	Location of the file
`RunNum`	Run number of the result
`PreProcParamList`	Parameter list for preprocessing
`PostProcParamList`	Parameter list for postprocessing
`ARisPrivate`	Privacy of the micrograph

The contents of the benchmark suite can be visualized in a high dimensional space where each axis represents a property of the micrographs or performance results of the various algorithms. Then for a given micrograph, a researcher would manually be able to pick the best particle identification method even with a limited knowledge of the properties of the micrograph e.g., noise parameters and CTF parameters. With proper metrics, this task may also be automated. As suggested in [25], two sets of results may be kept; one for the results on a set of freely available set of micrographs and the other for a set of micrographs that are not revealed. The database in the image repository contains a collection of tables. A non exhaustive list of these tables along with their corresponding fields and their interrelationships is shown in Figure 5.2.

A schematic of the design of a virtual workbench is shown in Figure 5.3. A description of the components of the workbench follows –

1. Client – The client is a **Java**[2] based applet which provides a secure access to the benchmark repository. Users can send requests for viewing images and results, uploading an image, and uploading or updating the results obtained on images in the repository by algorithms submitted by them. Access to the repository is granted by an administrator by creating an account that the user requests. When uploading micrographs, the users have to provide, in addition to the micrograph image, the information related to the micrograph such as those mentioned in table 5.1 and a brief description of macromolecule including keywords uniquely describing the macromolecule. When submitting the results, the user provides the information mentioned in 5.2. The users also have the choice of submitting results of running the algorithms they have submitted to the workbench on the micrographs

[2] **Java**™ is a registered trademark.

Figure 5.1: Design schematic of the internal of the Workbench application server

present in the workbench. These algorithms may be executed offline at the user's site. The client also enables collaboration between the users by allowing them to extend full/limited access to their micrographs, algorithms and results to selected workbench account holders. Once the access is granted these other members then have the ability to download those micrographs, algorithm code/executables/libraries and/or run the algorithms remotely at the workbench site.

2. Application Server – The server manages the user request for uploading image and results of algorithms on images in the repository. When a request for uploading an image, the request is received by the server, the image is stored in the repository and a unique ID for the image is created and the image is marked as 'not processed' i.e. *IsDirty* (see table 5.1) is set to true. When a request for updating the result for a specific image is received, the image is marked as 'new result in repository' and the result specific parameters are updated in the repository.

3. Image Upload Watcher (IUW)– The image upload watcher periodically looks into the repository for all the images that have been marked as 'not processed'. Based on the resource availability and the number of images to be processed, it spawns metadata extractors (ME) for extracting image parameters. There are certain limitations that apply on the micrographs before they can be automatically processed for metadata extraction e.g. the availability of 'ground truth' before the CTF noise and background parameters can be extracted.

4. Metadata Extractor (ME) – The metadata extractor takes a single image from the repository as indicated by the IUW and computes the relevant image parameters for it's characterization mentioned in Chapter 6. It then

updates the repository with the computed parameters corresponding to the image and marks it as 'processed'. Normally, each image goes through this process only once. However the arbitrator may choose to run them more than once.

5. Benchmark Repository Watcher (BRW) – The benchmark repository watcher periodically looks for all the images that have been marked as 'processed' or 'new result in repository' and sends a formatted message (email) to the arbitrator for review.

6. Arbitrator – The arbitrator is a human involved in the loop. Upon receiving a message (e.g. email) from the BRW, the arbitrator verifies all the information related to the uploaded micrograph images, algorithms and/or results of algorithms, and marks them as 'clean'/'invalid'. If any of the data if deemed invalid by the arbitrator, the client who uploaded that data is informed of an invalid data upload.

7. Benchmark Repository (BR) – The benchmark repository is a database of metadata for micrograph , metadata for algorithms (code, executables and libraries), and metadata for the results of the algorithms on micrographs contained in the database. The micrograph images, algorithms (code, executables and libraries) and results of the algorithms are stored in the file system.

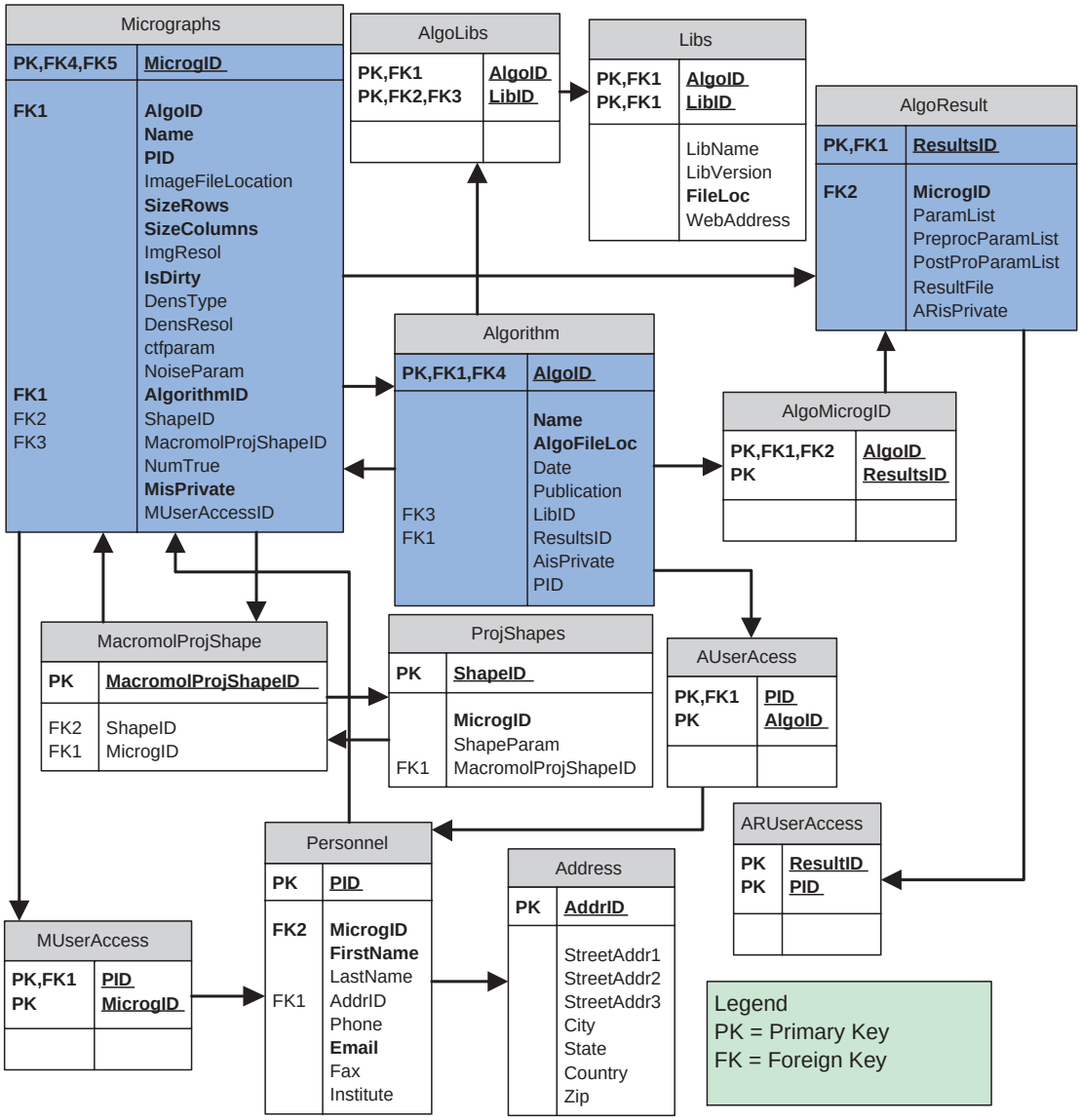


Figure 5.2: A schematic of the database tables and their interrelationships in the virtual workbench. The fields shown are a subset of the list of fields present in each of the tables.

5.4 Implementation

The implementation of each of the components described in the section 5.3 is given below. All the components of the system have been built using open source software.

1. Client – The clients have been implemented using the **Java**. The client interface is further described in section 5.4.1.

2. Server – The server consists of a web server and an application server. The web server provides a web access to the users. The application server provides a mechanism for the user to execute tasks on their behalf on the application server. The application server may spawn tasks on other systems as shown in Figure 5.3. The open source application server Apache[3] Tomcat is used as the application server. It hosts the Java servlets[4] which are used to allow clients to remotely execute programs on the webserver and the execution engines. The webserver integrated with the application server serves the HTML pages to the users. The execution engines are back-end machines that are used to run the algorithms and perform micrograph characterization. These machines are accessible using the executables run by the application server using the servlets as shown in Figure 5.4.

3. Image Upload Watcher (IUW) – The image upload watcher is implemented using `Perl`[5].

4. Metadata Extractor (ME) – The ME is written in `Perl`. The procedures for computing the image statistics are implemented in C. The ME is a col-

[3] **Apache©**
[4] **Servlets™**
[5] **Perl©**

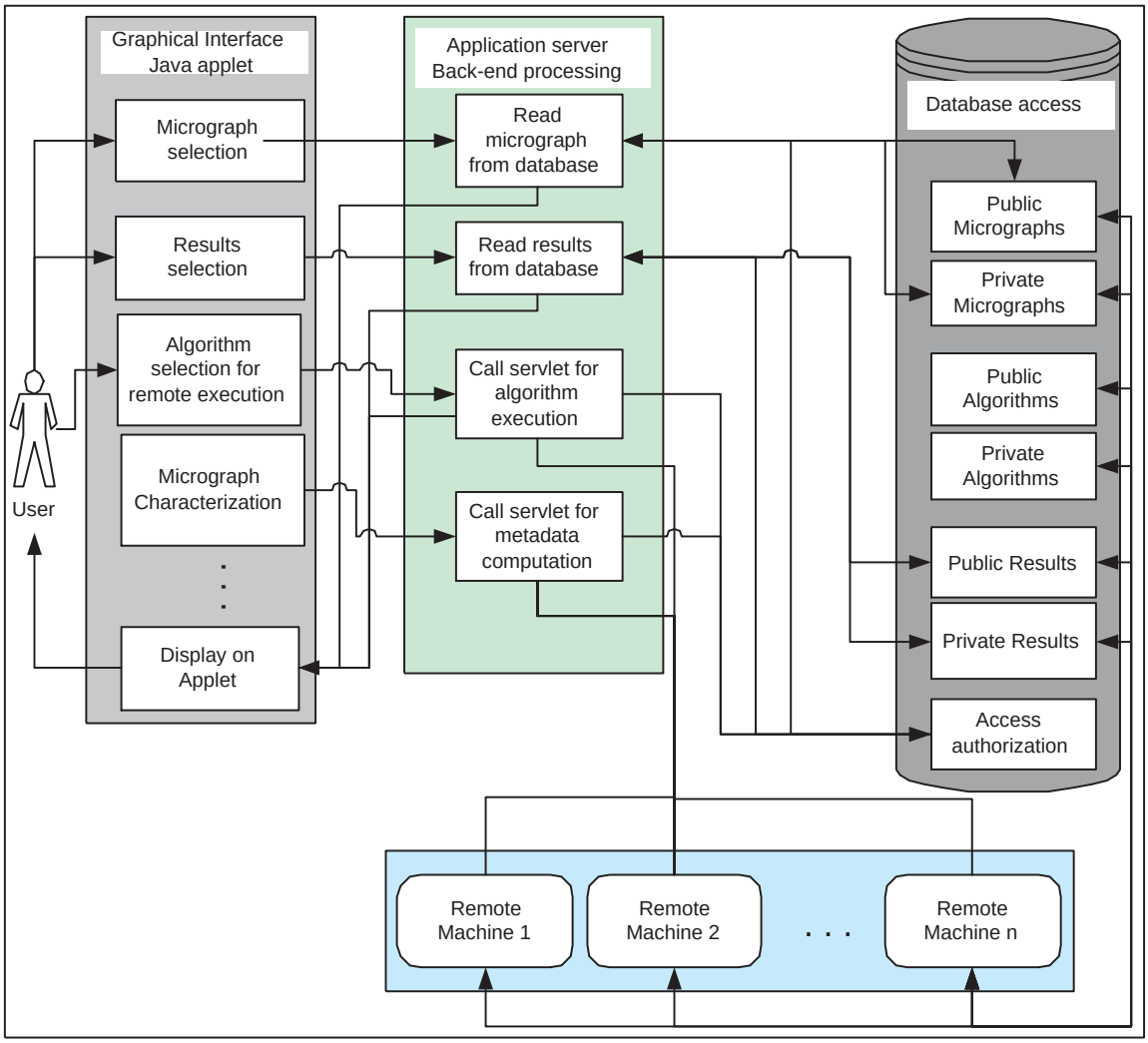


Figure 5.3: Data flow in the virtual workbench.

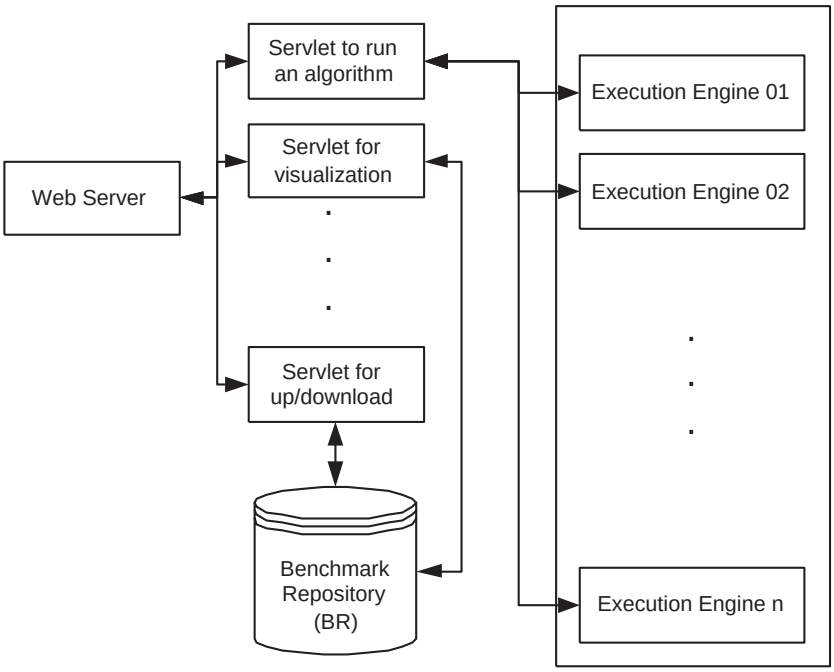


Figure 5.4: The interaction of servlets with other components in the workbench.

lection of calls to the C programs for computation of image statistics and database method calls for insertion of the image parameters in the database. For instance, the field *MaxBackApproxIntensityVar* is the maximum variation in the background of the image and is computed by modeling the background as a smoothing spline and computing the difference of the maximum and minimum values of the function over the image. Such computations are done using C. Invocation of the different C programs is done through the servlets. After all the image parameters have been computed, the *ParamsComputed* field of the record corresponding to the image is set to true.The Metadata Extractor is a collection of calls to the C programs for computation of image statistics and database method calls.

5. Benchmark Repository Watcher (BRW) – The benchmark repository watcher is a Java servlet which is executed whenever any data is uploaded.

6. Arbitrator – The arbitrator is a human involved in the loop.

7. Benchmark Repository (BR) – The open source database PostgreSQL[6] is used as the back-end data repository. The design of the database is depicted in Figure 5.2.

5.4.1 Client Interface

The users access the workbench using the client which is the only point of interface between the users and the workbench. The client is implemented as a Java applet which allows the user to access the workbench using any Java enabled web browser. The applet allows two levels of access to the users –

1. Semi-privileged access

2. Privileged access

The unprivileged level is the default access level. At this level of access, the users are allowed to view and download micrographs, algorithms and algorithm results that are available for public access. The privilege level is obtained by requesting an account. The arbitrator creates the account on behalf of the user requesting the account. The privileged level of access gives the users, in addition to the services given at the unprivileged level, the following services –

1. Upload/Download of micrographs, algorithms and algorithm results.

2. Remote execution of algorithms.

[6] **PostgreSQL©**

3. Characterization of private micrographs.

4. Maintain private database of micrographs, algorithms and algorithm results for a limited time.

5. Extend access to private micrographs, algorithms and algorithm results to other selected privileged users.

The main panel of the Java applet is shown in figure 5.5. The set of controls marked (A) allow the user to connect to different databases, perform account management for their account and access the help system of the applet. The databases may be located at geographically different sites. The set of controls marked (B) allow the user to explore the micrographs present in the workbench, upload and download micrographs and perform characterization of the micrographs uploaded. The set of controls marked (C) allow the user to access the algorithms and results of the algorithms on micrographs in the workbench, visualize the performance of the algorithms, and upload and download results of the algorithms. The controls in (D) allow the user to control the zoom, contrast and brightness level in the micrograph display. A thumbnail view of the micrograph currently selected is shown in (E). Further details about the panels is shown in Figures 5.6 and 5.7.

5.5 Conclusion

The virtual workbench is a tool to enable benchmarking current and future particle identification algorithms and promote collaboration between researchers for the further development of automatic particle identification algorithms. The

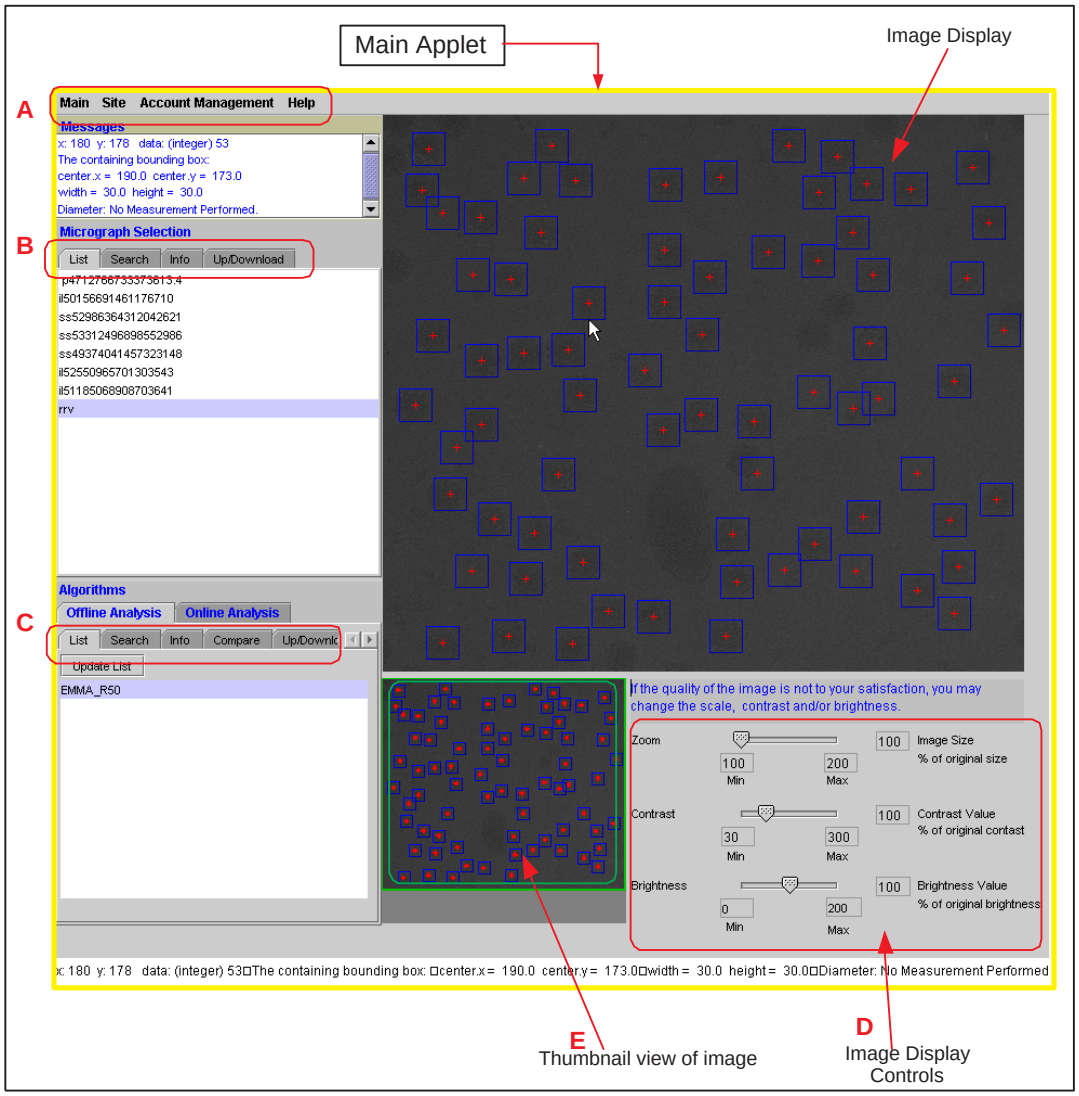


Figure 5.5: The main applet view is shown (bounded by yellow box) with views of the panel. The applet allows the user to control the zoom, contrast and brightness levels of the image displayed. A thumbnail view of the micrograph is provided.

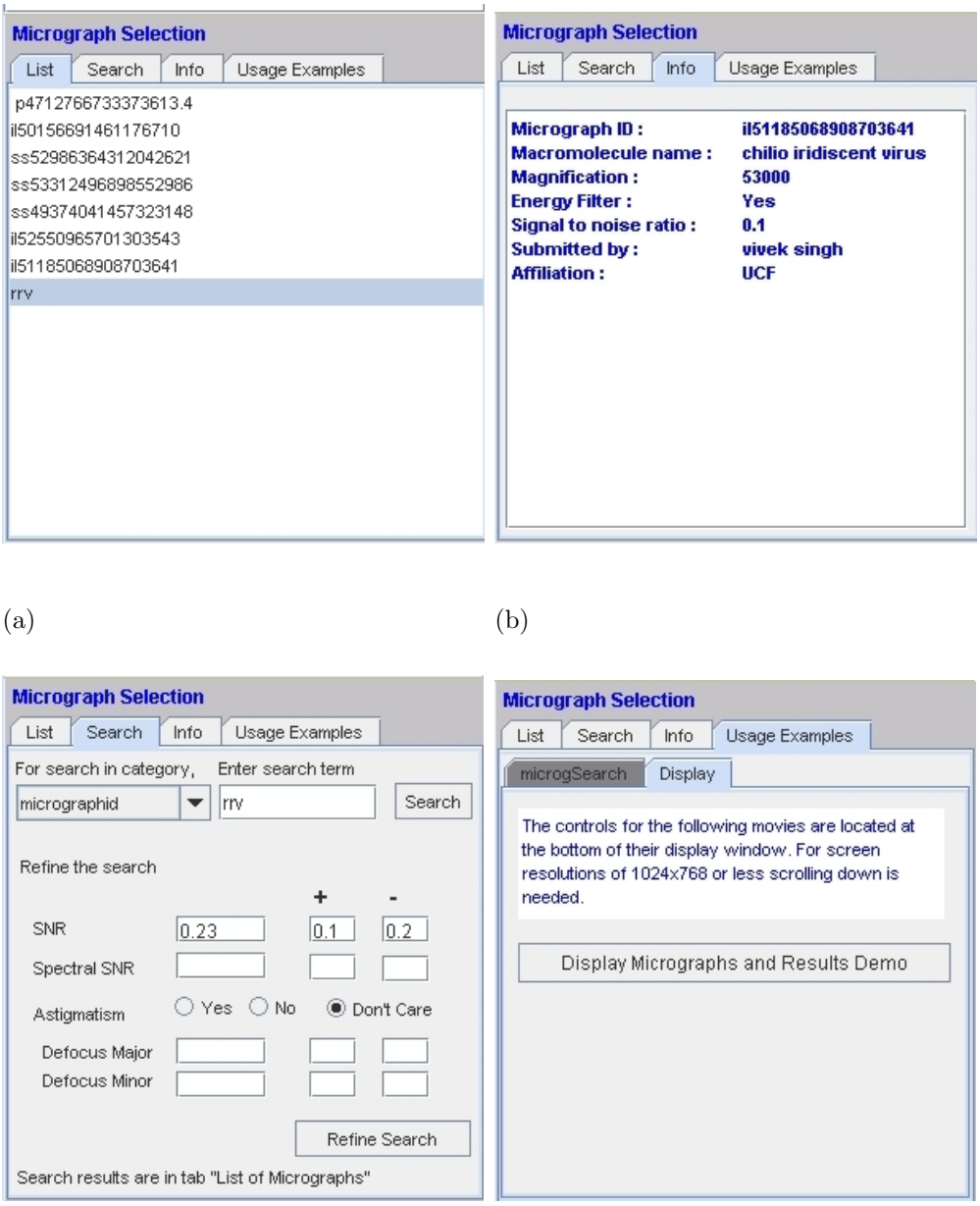


Figure 5.6: Various applet for micrographs are shown. (a) list of micrographs in the workbench, (b) view information about the micrographs, (c) search for micrographs based on various fields, and (d) usage information.

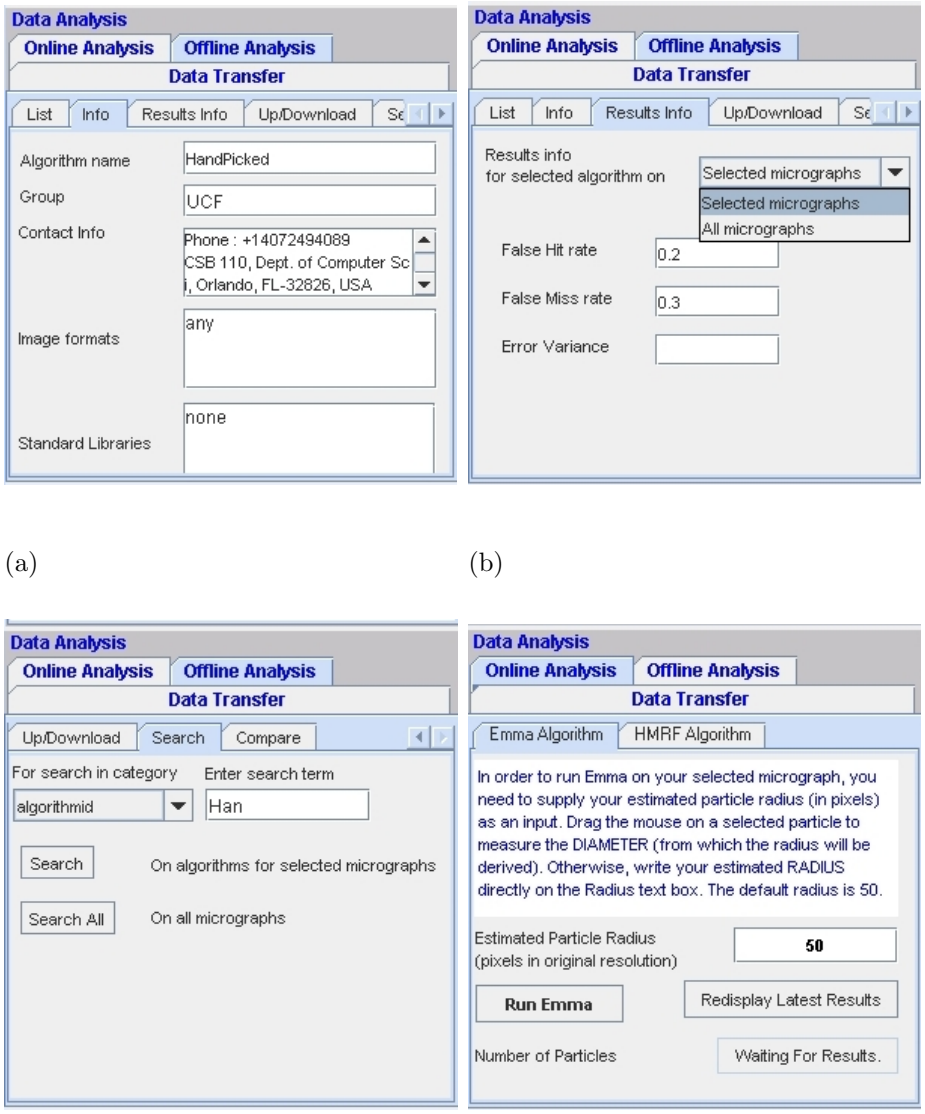


Figure 5.7: Various applet panels for algorithms are shown. (a) List of algorithms that have results for the selected micrograph/s in the workbench, (b)view information about the algorithms, (c) view information about the results of the selected algorithm on the selected micrograph/s (d) search for algorithms based on various fields, and (e) remotely execute a selected algorithm on the workbench for selected micrograph.

workbench is web accessible and can be accessed using any browser enabled with Java™. The system is open and the users of the system contribute to it's content. Users have an ability to upload both micrographs, particle identification algorithms and results of the selection of the particle identification algorithms. The virtual workbench can be easily extended to address similar interests in the single particle three-dimensional reconstruction research.

CHAPTER 6

CHARACTERIZING MICROGRAPHS AND ALGORITHMS

No amount of experimentation can ever prove me right; a single experiment can prove me wrong.

— *Albert Einstein.*

An essential part of the virtual workbench is the quantitative characterization of micrographs and the particle identification algorithms. Such quantitative characterization facilitate comparisons and benchmarking. They also allow the users to better understand the performance of the algorithms with variation in the 'types' of the micrographs.

6.1 Characterization of Micrographs

A characterization of the micrographs is an important aspect of the Virtual workbench if one wishes to explore the variation of the 'performance' of particle identification algorithms with the different 'types' of micrograph. Within the context of such a characterization, a micrograph must be associated, among other things, with a set of parameters that are closely associated with the imaging process. Such parameters include the defocus level, lens current etc. There are some other parameters that must be obtained by processing of the micrographs after the imaging process is completed. We discuss some of these parameters in the following.

The micrographs can be characterized by the following non exhaustive list of parameters that are obtained by processing of the images.

1. Type of noise and it's parameters.
2. Signal to noise ratio.
3. Contrast transfer function parameters.
4. Background.

In addition to the parameters mentioned above, the type/s of macromolecule particle whose projections are captured by the micrograph is also used to characterize the micrographs. The workbench also contains ancillary data associated with each micrograph regarding the personnel involved in the imaging of the micrograph, the date the images were captured and the group involved.

Many of the particle identification algorithms do not vary significantly in their performance along certain parameters. However, due to the fact that the virtual

workbench is an open system, algorithms added to it in the future may exhibit a variation in their performance along those parameters. Hence although some of the parameters do not seem to differentiate between the particle identification algorithms, they are still kept for characterization because of their saliency. In the following, each parameter is described.

6.1.1 Particle type

The type of symmetry of the particles, and more precisely, that of the projections sometimes affects the performance of some particle identification algorithms that are designed for a particular type of projection shapes [43], [70], [90], and [94]. Hence the particle projection shapes corresponding to each micrograph is maintained. In some cases where a micrograph has a collection of projections with more than one shapes, the micrograph may be associated with more than one shape corresponding to the different projections [92] whether the projections correspond to the same type of particle or not.

6.1.2 Noise and its parameters

The performance of particle identification algorithms depends, among other things, on the noise characteristics of the micrographs. An estimation of the noise characteristics helps explore the "performance" of the particle identification algorithms with the "quality" of the micrographs. An assumption of an independent and identically distributed (iid) gaussian noise has been made for the micrographs . It has been validated in Figure. 3.1 and Figure 3.2. With this assumption of a

gaussian noise, the parameters of the noise namely the mean and variance are used to characterize a micrograph. The noise is not always spatially stationary e.g as shown in Figure 6.1 the mean of the background pixel intensity varies along the micrograph. Also, the variation of parameters over the same micrograph may vary for different micrographs. Hence an average of the parameters is used to characterize the noise content of the micrographs.

The parameters of the gaussian distribution can be easily estimated by computing the sample mean and variance. With the background regions known, sections of the micrograph are extracted that correspond to the background. Sample means and variances for each region is computed individually. The mean of these sample means and variance is computed and stored as the noise parameters of the gaussian noise in the micrograph.

6.1.3 Signal-to-noise ratio

For a given image, the signal-to-noise ratio can be defined as the variance of the signal to the variance of the noise[3].

For an image M^ℓ, the sample variance is defined as

$$var(M^\ell) = \frac{1}{S-1} \sum_{i=1,j=1}^{m,n} [M^\ell_{i,j} - \bar{M^\ell}]^2$$

where,

$$\bar{M^\ell} \;=\; \frac{1}{S} \sum_{i=1,j=1}^{m,n} M^\ell_{i,j}$$

[3]The signal-to-noise ratio is sometimes defined as $\frac{Power\ of\ the\ signal}{Power\ of\ the\ noise}$. However here we use the former definition

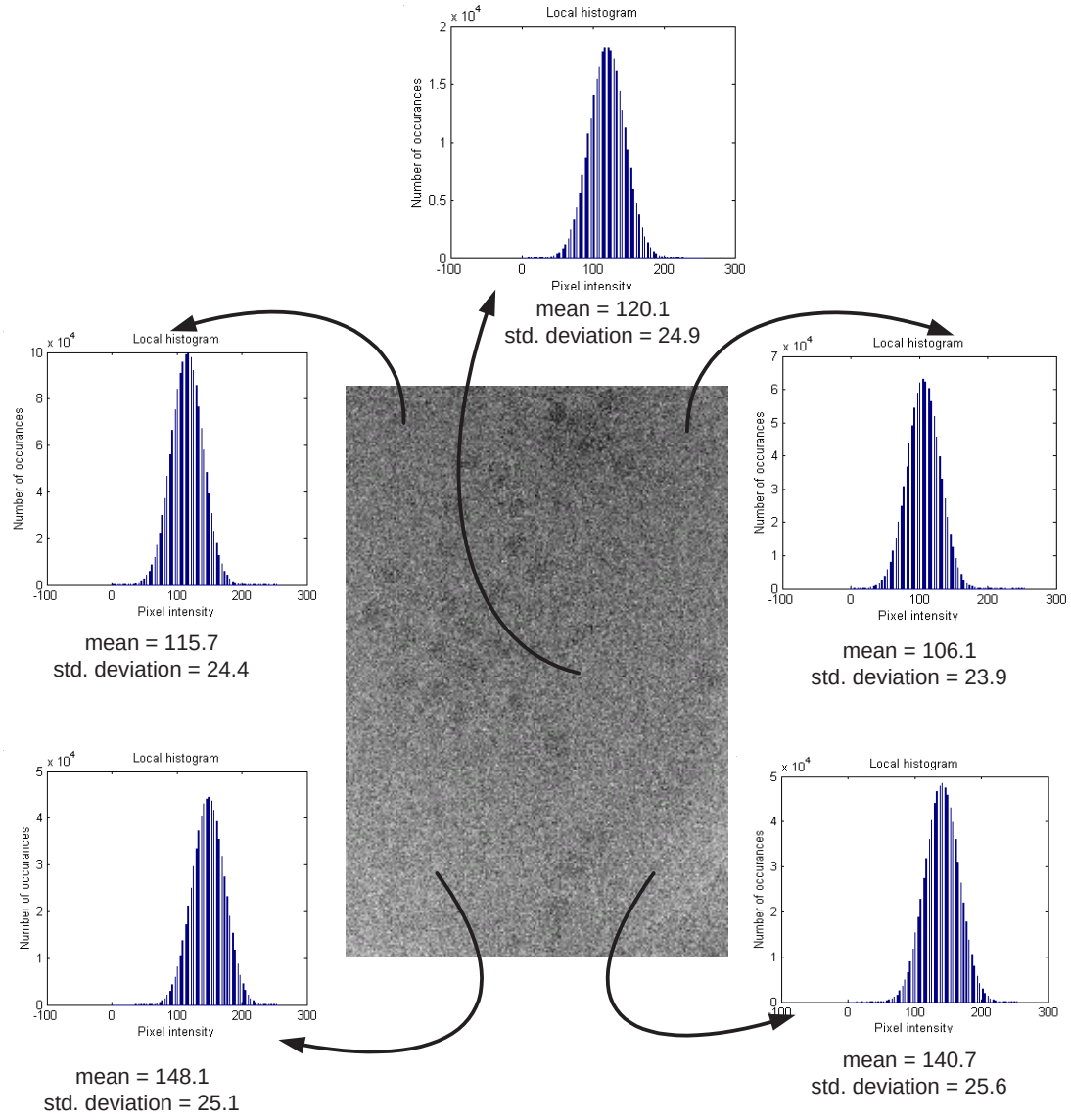


Figure 6.1: The variation of the mean intensity of the background pixels across the micrograph image. The mean intensity of pixels within the projections of virus have a similar variation.

Measurement of the signal to noise ratio is a difficult task due to the variation in the background, the method proposed in [18] can be utilized. The method is based on computing the cross correlation coefficient of two images recorded for the same area. If the images have sufficiently large number of pixels, the SNR value can be computed using the cross correlation coefficient of the images. For the two images p_{ij} and p_{kj}, the cross correlation coefficient is given as

$$\rho_{i,k} = \frac{\sum_{j=1}^{J}[p_{ij} - < p_i >][p_{kj} - < p_k >]}{\sqrt{\sum_{j=1}^{J}[p_{ij} - < p_i >]^2 \sum_{j=1}^{J}[p_{kj} - < p_k >]^2}}$$

Although the condition of p_{ij} and p_{kj} being the images of the same region of the image is not satisfied as shown below, nevertheless we will use this method for computing the SNR values using the following scheme.

Given : Micrograph M, s true locations of the particle projections

- Compute the $C = sxs$ matrix where each entry (i,k) corresponds to the cross correlation coefficient of the images p_i and p_k.
- For each row of C, find the maximum entry.
- Use the entry in step 2 to compute the SNR at each row.
- Add the SNR values for all the rows to get S_{SNR}. Average SNR then becomes $s_{SNR}/(2 * s)$

6.1.4 Contrast transfer function

The estimation of contrast transfer function plays an important role in the $3D$ reconstruction of macromolecules within realm of cryoTEM. A brief description

of the contrast transfer function is given in section 2.1.1.3 with further details given below.

A micrograph image results from a combination of sample-induced elastic and inelastic electron scattering. In general, inelastic scattering produces an almost featureless background in the image power spectrum that is high at low resolution and falls off toward higher resolution. The amplitude contrast in micrograph mainly arises from the removal of high-angle elastic scattered electrons that are outside the objective aperture. The elastically scattered electrons that pass through the objective aperture produce the phase contrast that contains most of the structural information in cryoTEM of specimens. The amplitude and phase contrast are modulated by the CTF of the microscope which is a function of defocus, astigmatism, lens errors, electron wavelength, as well as temporal and spatial coherence of the electron beam [95]. In general, the average of the power spectra obtained from individual molecular images in one micrograph (which have a nearly identical CTF) is used to determine the CTF parameters. Although a precise determination of CTF based solely on the shape and position of the Thon rings (or ellipses) is difficult as it would require knowledge of all parameters and a better understanding of inelastic and multiple scattering, the oscillations due to the CTF can be interpreted in a straight-forward manner. The phase reversals can be roughly described by a simple oscillation function of constant amplitude, determined by factors such as the defocus, astigmatism and spherical aberration of the objective lens. A more accurate description requires this periodic function to be multiplied by an envelope function which monotonically attenuates the CTF toward higher spatial resolution and captures the effects of spatial and temporal beam incoherence. The expression for the two-dimensional CTF is given by –

$$CTF \;=\; \sqrt{1 - A^2} * sin(\theta) + A * cos(\theta)$$

where

$\theta = 2\pi * (\frac{\Delta f * \lambda * d^2}{2} - \frac{C_s * \lambda^3 * d^4}{4})$.

A - The fractional amplitude contrast.

λ - The electron beam wavelength.

d - The spatial frequency related to the coordinate (h, k) in the Fourier domain.

C_s - A spherical aberration constant of the microscope's objective lens.

Δf - The defocus is a function of minimum defocus, D_a, maximum defocus, D_b, and the angle α of the main axis of the ellipse with the h axis as shown in figure 6.2

$\Delta f = D_a cos^2(\phi - \alpha) + D_b sin^2(\phi - \alpha)$

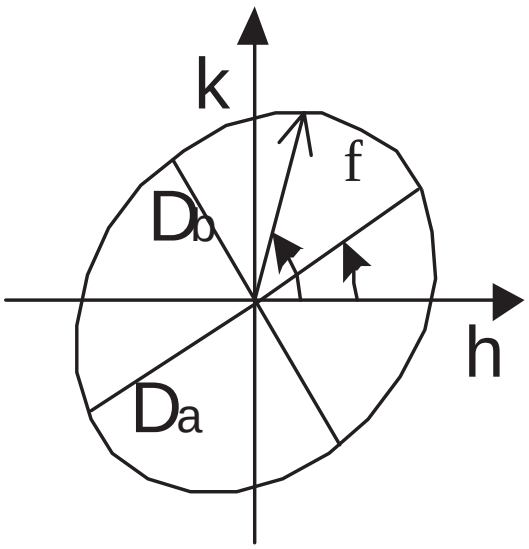


Figure 6.2: Parameters needed for CTF computation.

To account for the spatial and temporal coherence, the CTF is then multiplied by two separate exponential decay functions Es and Et [91]: $CTF = CTF * Es * Et$. In our algorithms we apply the CTF correction and a single attenuation correction factor, the inverse Temperature Factor (TF) to the Discrete Fourier

Transform of each raw image before reconstruction:

$$FFT_new(h,k) = \frac{FFT_old(h,k)}{(CTF(h,k) + WienerLike)} * \frac{1}{TF(h,k)}.$$

Our algorithm determines the defocus minor, D_a, the defocus major, D_b, the angle α , and a temperature factor that takes into account several effects which lead to an attenuation of the CTF. We compute a two-dimensional function to fit the oscillations of the average power spectrum in two dimensions. The correlation of the average power spectrum with a CTF pattern allows us to determine the level of astigmatism and the temperature factor. Following is a brief sketch of the algorithm –

Algorithm: CTF computation

1. Read the input data
   (a) Read size of the DFT, pixelsize, A, C_s, voltage, image data file;
   (b) Read initial estimates of: D_a, D_b, α , and the temperature factor;
2. Read each boxed image;
3. Calculate the 2D-DFT of each image; sum the 2D-DFTs together; calculate the unweighted average of the DFT, the compute the modulus of the average value (DFT_{avg})
4. Determine the angle using a multi-resolution refinement procedure;
   (a) For a particular refinement step size (say 1) rotate clockwise DFT_{avg} and correlate the upper and the lower half, of DFT_{avg} relative to the h axis. Determine the angle α which maximizes the correlation coefficient;

(b) Refine the angle (say to 0.1, then 0.01) and continue the search around the value obtained in 4a;

5. Estimate a background function and subtract the background from DFT_{avg}.

6. Determine the defocus major, the defocus minor, and the temperature factor using the initial input parameters:

   (a) For a given step size, resolution, and window size, create a set of CTF patterns in the range of (D_a ± w*r, D_b ± w*r, temperature factor ± w*r), with w - the window size, the range of the defocus major and minor and the temperature factor.

   (b) Compute the correlation coefficient of the DFT_{avg} with these CTF patterns. Find the maximum correlation coefficient and the corresponding values for: D_a, D_b, and the temperature factor.

   (c) If needed, slide the window w, and repeat steps 6a and 6b.

   (d) Decrease the step size, repeat steps 6a - 6c and refine defocus major, defocus minor, temperature factor.

7. Verify the convergence of the algorithm using a different set of input parameters.

The algorithm was tested with 198 particle projections from one micrograph, each image being of size 255x255 pixels. Figure 6.3 (a) shows the DFT_{avg} of these images. We started with an initial guess: $D_a = 2.0\mu m$, $D_b = 2.0\mu m$, $\alpha = 0.0°$, temperature factor = 100.0 Å^2. The parameters produced by the algorithm: $D_a = 2.503\mu m$, $D_b = 3.288\mu m$, $\alpha = 7.89°$, temperature factor = 723.0 Å^2. Figure 6.3 (b) shows a superposition of the power spectrum and the CTF function of Figure 6.3 (a). A match in the Thon rings with the ellipses gives a qualitative

indication that the algorithm works well. We are also encouraged because: (a) the correlation coefficient for the optimal set of parameters is better than of 0.99, and (b) starting with a different set of initial guesses we reached the same optimal parameter values.

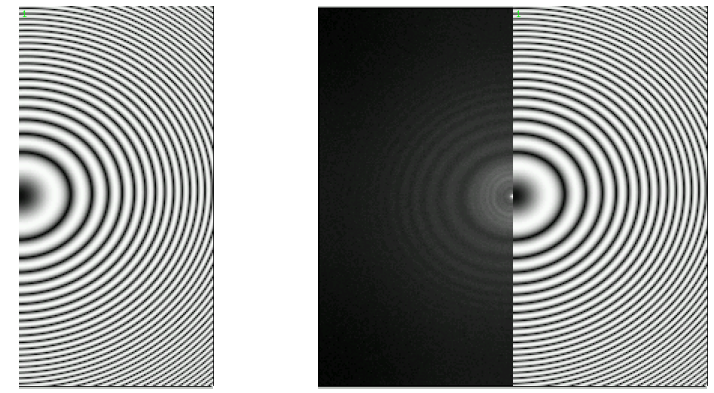


Figure 6.3: For an average CTF computed for 198 particle projections, (a) shows the computed CTF and (b) shows that the Thon rings/ellipses match when computed CTF is superposed with the average CTF that was used to compute the CTF parameters

6.1.5 Image background

Cryo TEM micrographs generally contain, in addition to the noisy projections of the imaged particles, some artifacts due to ice formation and other impurities. The image scanning process also contributes to the creation of artifacts. Besides these, there is a presence of a gradient in the background. An estimation of this gradient has been used to characterize micrographs. A quantitative measure of the background must be computed if we hope to use it as a parameter for micro-

graph characterization. We do this by fitting a thin plate spline to a sub-sampled version of the micrograph with projections and artifacts due to ice and impurities removed. The procedure is shown below along with a schematic illustrated in figure 6.4.

1. From the given ground truth data, those portions of the micrograph that correspond to particle projections, and artifacts due to impurities and ice formation are removed.

2. The remaining micrograph is divided into contiguous blocks. The mean of each block is computed.

3. A thin plate spline is fitted on a two dimensional array of these mean values.

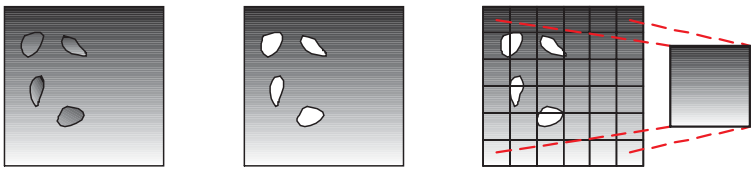


Figure 6.4: Schematic for background estimation.

The procedure can be used for micrographs without the ground truth data. Such micrographs are run through the particle identification algorithms and all the regions that are classified as particle projections are removed from the micrograph before computing the background. However, such an automatic method is closely tied to the automatic method used for identifying the projections. A consensus based approach can also be taken where a collection of identification algorithms are allowed to identify the particles and those particles that are voted more than a threshold level are considered to be the valid projections and the background is then computed after removing them from the micrograph image.

6.2 Characterization of Algorithms

At the fundamental level, the aim of any particle identification algorithm is the identification of the locations on the micrograph where projections of the particle/s under consideration are present. For micrographs containing projections of only a single type of particle, the locations of projections of the particle on each micrograph is sufficient. However when the projections due to more than one type of particle are present on the micrographs, the type of particle that each projection represents must also be determined by the particle identification algorithm.

6.2.1 Performance metrics

The performance of particle identification algorithms is traditionally expressed as the false hit rate and false miss rate, definitions of which are given below. Manual identification of particle projections by a human expert is always considered as the ground truth. Let us assume that for a given micrograph M^{ℓ} (where ℓ ranges over the collection of micrographs, in our case in the virtual workbench, that have the ground truth available for them) containing projection of $i = 1$ to N particles, the following is the ground truth.

1. The number of particle projections : $\hat{k}_0^{\ell}$,
2. The estimated location of particle projections :
   $\{(x_1^{0,\ell}, y_1^{0,\ell}), (x_2^{0,\ell}, y_2^{0,\ell}), ...(x_{\hat{k}}^{0,\ell}, y_{\hat{k}}^{0,,\ell})\}$, and
3. The type of particle for each projection : $\{t_1^{0,\ell}, t_2^{0,\ell}, ...t_{\hat{k}}^{0,\ell}\}$.
   where $t_j^{0,\ell} = \ i, \ j = 1 \ to \ \hat{k}_0^{\ell}$, and $\ i \in \{1, ..., N\}$.

Now in order to characterize a particle identification algorithm alg_A we need the following data for each micrograph M^ℓ.

1. The number of particles are identified : $\hat{k_A^\ell}$,
2. The estimated location of particle projections :
   $\{(x_1^{A,\ell}, y_1^{A,\ell}), (x_2^{A,\ell}, y_2^{A,\ell}), ...(x_{\hat{k_A^\ell}}^{A,\ell}, y_{\hat{k_A^\ell}}^{A,\ell})\}$, and
3. The type of particle for each projection : $\{t_1^{A,\ell}, t_2^{A,\ell}, ...t_{\hat{k_A^\ell}}^{A,\ell}\}$.
   where $t_j^{A,\ell} = \ i, \ j = 1 \ to \ \hat{k_A^\ell}$, and $\ i \in \{1, ..., N\}$.

For fixed values of tolerances τ_x along the x axis and τ_y along the y axis, the *false positives rate* anf *false negative rate* are defined as follows –

- *False positive rate* A false positive event occurs when the algorithm "discovers" false or unwanted particles ("junk"). Quite often the centers of the particles are used to identify its location. However, when the particle projection is not circular, then either the center of mass can be used or one of the predefined corners e.g. 'top-left' of the bounding box will suffice as the position of the projection. Number of positions on the micrograph where the algorithm indicates the presence of a projection when there is none at that position within a tolerance of τ_x along x axis and τ_y along y axis is the number of false positive events.

  Mathematically, the number of false positive events is given by

  $$FPE = \sum_{\ell} \sum_{p=1}^{\hat{k_A^\ell}} \ Pos_p^\ell$$

  where Pos_p^ℓ is defined as,

$$Pos_p^\ell = \begin{cases} 0 & \text{if } (|x_p^{A,\ell} - x_q^{0,\ell}| \leq \tau_x) \; and \; (|y_p^{A,\ell} - y_q^{0,\ell}| \leq \tau_y) \text{ for some } q \in \{1...\hat{k^\ell}\} \\ 1 & \text{otherwise} \end{cases}$$

FPE is the number of false positive events for the algorithm alg_A on the set of micrographs M^ℓ. The *false positive rate* (FPR) given by

$$FPR = \frac{FPE}{\sum_\ell \hat{k^\ell}}$$

- *False negative rate* A false negative event occurs when the algorithm fails to identify genuine particle projections. Number of positions on the micrograph where the algorithm does not indicate the presence of a projection within a tolerance of τ_x along x axis and τ_y along y axis when there is a projection at that position as per the ground truth is the number of false negative events. The concept of tolerance is given in figure 6.5.

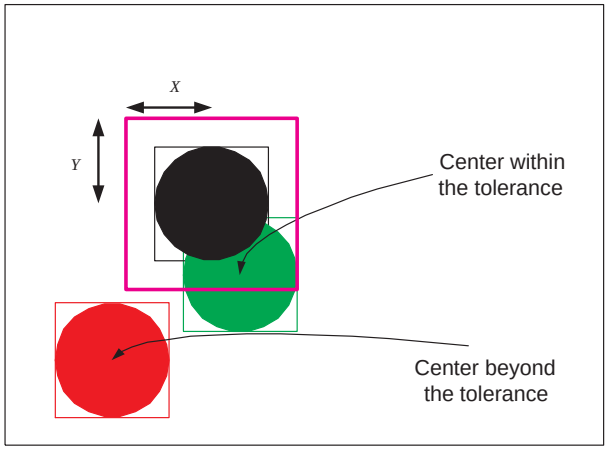


Figure 6.5: An illustration of the concept of tolerance. The label X refers to τ_x and Y refers to τ_y. These tolerances can be set by the user.

Mathematically, the number of false negative events is given by

$$FNE = \sum_\ell \sum_{p=1}^{\hat{k_A^\ell}} Neg_p^\ell$$

where Neg_p^ℓ is defined as,

$$Neg_p^\ell = \begin{cases} 1 & \text{if } (|x_p^{A,\ell} - x_q^{0,\ell}| > \tau_x) \; and \; (|y_p^{A,\ell} - y_q^{0,\ell}| > \tau_y) \text{ for all } q \in \{1...\hat{k^\ell}\} \\ 0 & \text{otherwise} \end{cases}$$

FNE is the number of false positive events for the algorithm alg_A on the set of micrographs M^ℓ. The *false negative rate* (FNR) given by

$$FNR \; = \; \frac{FPE}{\sum_\ell \hat{k^\ell}}$$

Although the aforementioned metrics are most often used, some new metrics can be defined as shown below.

1. Mean square error in estimating a location ($\bar{E_A}$)and
2. Variance of the error ($Var(E_A)$).

Here the error is the Euclidean distance between the location of a projection as obtained by an algorithm and the true location (e.g. center of the projection) as given by the ground truth data. Care must be taken to disregard those projections that are at a distance of more than τ_x and τ_y from true locations.

6.2.2 Synthetic data

If the knowledge of the data generation process is available, the use of synthetic data can be made for testing algorithms. Even in the case of a complete knowledge not being available for the data generation process, the performance of the algorithms over synthetic data can give insights towards the working of the algorithms. Within the context of cryo TEM, the data generated through the imaging

process can be approximated to generate synthetic micrographs. We start with a three dimensional model of a macromolecule $V(i, j, k)$, where i, j, and k vary from 1 to s along, say a cartesian axis thus sampling the structure V at r^3 points in space. The sampling may also be done in polar coordinates or cylindrical coordinates with i, j and k appropriately defined.

Using the given three dimensional model, radon transform is utilized to get the projections p_i of the model along different directions. A collection of such projections is given in figure 6.6.

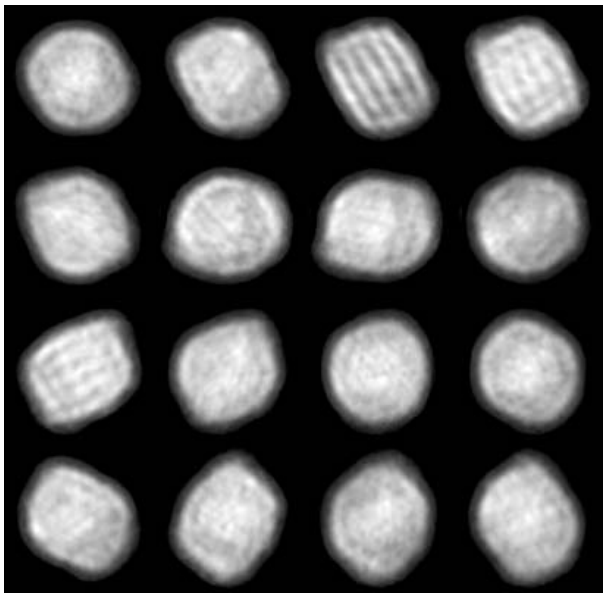


Figure 6.6: Radon transform of a $3D$ model along a set of 16 lines. The radon transform is computed along 16 lines that pass through the center of the $3D$ model. The direction of the lines is randomly distributed

The steps involved in creating a synthetic micrograph is given below –

- For each projection p_i, it's Fourier transform is performed to obtain P_i, $i \in 1...s$

- P_i is multiplied by the contrast transfer function at some set parameters to get Q_i
- Inverse Fourier transform is performed on Q_i to get q_i.
- Each q_i is rotated at a random angle and distributed over the image plane at random locations.
- Gaussian noise is added to the image to get the synthetic micrograph.

6.2.3 Analysis of performance

Analysis of the particle identification algorithms must give the users an insight into the variation of performance of the algorithms with variation in values of the parameters characterizing the micrographs. The users must be able to judge the those parameters of the micrograph characterization that have a more prominent impact on the performance of the identification algorithms than those that have a less prominent impact.

In order to perform a qualitative analysis of the performance of the particle identification algorithms, the following is done. For each of the four parameters of the performance characterization of the algorithms viz. false positive rate, false negative rate, mean square error and error variance, an individual plot is generated against each parameter characterizing the micrographs. The user can then view the plots and judge the parameters of micrograph characterization that have the most affect on the performance of the identification algorithms.

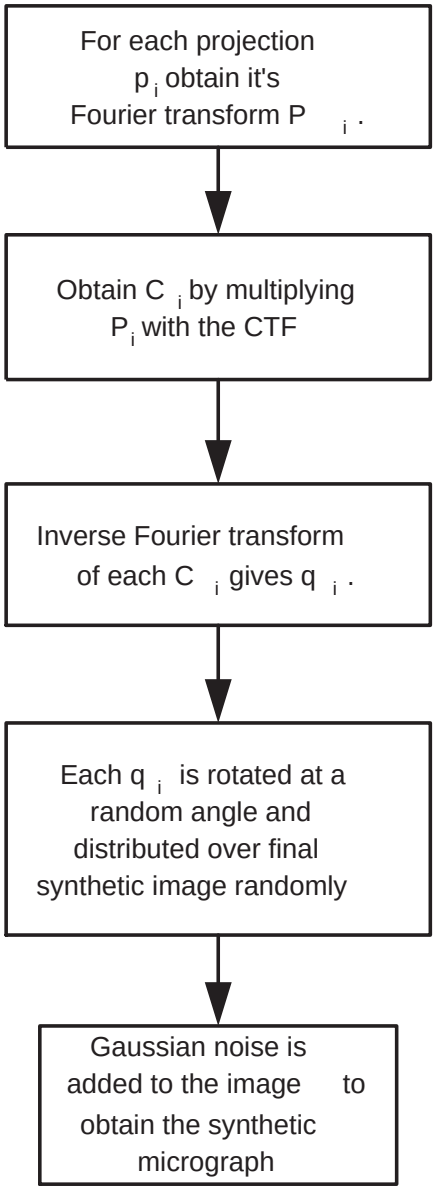


Figure 6.7: A scheme for generation of synthetic micrographs.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

An education isn't how much you have committed to memory, or even how much you know. It's being able to differentiate between what you do know and what you don't.

— *Anatole France.*

A summary of the main ideas developed is presented in the following section. Furthermore the future work is described along with sketches of possible approaches that might be taken.

7.1 Conclusions

A procedure based on the framework of hidden Markov random fields was developed for segmentation of micrographs. The anisotropic diffusion technique is used to pre-process the micrographs. The hidden Markov random fields procedure is based on a neighborhood of first order and it employs the expectation maximization algorithm. The optimization is achieved using the iterated conditional modes algorithm. Any particle boxing method may be used on the segmented image.

The serial algorithm was parallelized in order to take advantage of the massive computational power afforded by cluster of workstations. The parallel algorithm was shown to perform better than the sequential version.

A web based open system called "Virtual workbench" was developed in order to facilitate comparisons and benchmarking of particle identification algorithms and to provide a collaborative environment for analysis of particle identification algorithms. A method based on splines for estimating the background of the micrographs was also developed. Furthermore, some metrics based on statistical estimation theory were introduced for algorithm characterization.

7.2 Future Work

There are three main focus areas along which the future work would be aligned. The first area would be focused around development of more sophisticated particle boxing algorithms. This would involve the use of the preliminary `3D` model of the macromolecule under consideration. The second area would be focused around development of algorithms for automatic characterization the micrographs and algorithms using sophisticated statistical tools. This would entail better estimation of both the CTF and the noise parameters. The third area of focus would be towards extension of the workbench to three dimensional reconstruction algorithms.

LIST OF REFERENCES

[1] T. S. Baker and R. H. Cheng. A Model-Based Approach for Determining Orientations of Biological Macromolecules Imaged by Cryoelectron Microscopy, In *Journal of Structural Biology*, 116:120–130, 1996.

[2] T. S. Baker, N. H. Olson and S. D. Fuller. Adding the third dimension to virus life cycles: Three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs, In *Microbiology and Molecular Biology Reviews*, 63(4):862–922, 1999.

[3] D. Barash. A Fundamental relationship between bilateral filtering, adaptive smoothing and nonlinear diffusion equation, In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):844–847, 2002.

[4] D. M. Belnap, A. Kumar, J. T. Folk, T. J. Smith and T. S. Baker. Low-Resolution Density Maps from Atomic Models: How Stepping "Back" can be a step "Forward", In *Journal of Structural Biology*, 125:166-175,1999.

[5] J. Berriman and N. Unwin. Analysis of transient structures by cryo-microscopy combined with rapid mixing of spray droplets, In *Ultramicroscopy*, 56(4):241–252, 1994.

[6] J. Besag. On the statistical analysis of dirty pictures, In *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.

[7] M. J. Black, G. Sapiro, D. H. Marimont and D. Heeger. Robust Anisotropic Diffusion, In *IEEE transaction on Image Processing*, 7(3):421–432 1998.

[8] C. Bouman and M. Shapiro. A multiscale random field model for Bayesian image segmentation, In *IEEE Transactions on Image Processing*, 3(2):162–177, March 1994.

[9] J. Canny. A computational approach to edge detection, In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–697, 1986.

[10] B. Carragher, et. al. Leginon: An Automated System for Acquisition of Images from Vitreous Ice Specimens, In *Journal of Structural Biology*, 132(1):33-45, 2000.

[11] S. Chen, A. M. Roseman and H. R. Sabil. Electron Microscopy Chaperonins, In *Methods of Enzymology*, 290:242–253, 1998.

[12] R. A Crowther, D. J. DeRosier and A. Klug. The reconstruction of a three-dimensional structure from projections and its application to electron microscopy, In *Proceedings of the Royal Society of London*, A 317:319–340, 1970.

[13] A.P. Dempster, N.M. Laird and D.B. Rubin. Maximal likelihood form incomplete data via the EM algorithm, In *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

[14] J. H. Elder and S. W. Zucker. Local scale control for edge detection and blur estimation, In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7):699–716, 1998.

[15] J. J. Fernndez, J. R. Sanjurjo and J. M. Carazo. A spectral estimation approach to contrast transfer function detection in electron microscopy, In *Ultramicroscopy*, 68(4):267-295, 1997.

[16] M. Flynn. Some Computer Organizations and Their Effectiveness, In *IEEE Transaction on Computers*, C-21(9):948-960, 1972.

[17] H. Foroosh, J. B. Zerubia and M. Berthod. Extension of phase correlation to sub pixel registration, In *IEEE Transaction for Image Processing*, 11(3):180-200, 2002.

[18] J. Frank and L. Al-Ali. Signal to noise ratio of electron micrographs obtained by cross-correlation, In *Nature*, 256:376–378, 1975.

[19] J. Frank and T. Wagenkknecht. Automatic selection of molecular images from electron micrographs, In *Ultramicroscopy*, 2(3):169–175, 1983-84.

[20] J. Frank. Three-Dimensional Electron Microscopy of Macromolecular Assemblies, Academic Press, 1996.

[21] J. Frank, M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj, and A. Leith. SPIDER and WEB: Processing and Visualization of Images in 3D Electron Microscopy and Related Fields, In *Journal of Structural Biology*, 116(1):190–199, 1996.

[22] A. S. Frangakis and R. Hegerl. Noise Reduction in Electron Tomographic Reconstructions Using Nonlinear Anisotropic Diffusion, In *Journal of Structural Biology*, 135:239–250,2001.

[23] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[24] R.M. Glaeser. Electron crystallography: Present excitement, a nod to the past, anticipating the future, In *Journal of Structural Biology*, 128:3–14, 1999.

[25] R.M. Glaeser. Historical background: why is it important to improve automated particle selection methods?, In *Journal of Structural Biology*, In Press, 2003.

[26] R. C. Gonzales and R. E. Woods. *Digital Image Processing*, Prentice Hall, 1996.

[27] G. Harauz and F. A. Lochovsky. Automatic selection of macromolecules from electron micrographs, In *Ultramicroscopy*, 31:333–344, 1989.

[28] G. Harauz and M. van Heel. Exact filters for general geometry three dimensional reconstruction, In *Optik*, 73:146–156, 1986.

[29] H. J. A. M. Heijmans. Morphological image operators, *Academic Press, Boston*, 1994.

[30] R. Henderson. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstrained biological molecules, In *Quarterly Reviews of Biophysics*, 28:171–193, 1995.

[31] J. S. Lee and K. Hoppel. Noise modeling and estimation of remotely sensed images, In *Proceedings International Geoscience and Remote Sensing, 1989*, 2:1005-1008, 1989.

[32] W. Jiang, M. L. Baker, Q. Wu, C. Bajaj and W. Chiu. Applications of bilateral denoising filter in biological electron microscopy, In *Journal of Structural Biology*, In Press.

[33] L. Joyeux and P. A. Penczek. Efficiency of 2D alignment methods, In *Ultramicroscopy*, 92:33-46, 2002.

[34] S. Kirkpatrick, C. D. Gelart Jr. and M. P. Vecchi. Optimization by simulated annealing, In *Science*, 31:671-680, 1983.

[35] T. Kivioja, J. Ravantti, A. Verkhovsky, E. Ukkonen, and D. Bamford. Local average intensity-based method for identifying spherical particles in electron micrographs, In *Journal of Structural Biology*, 131:126–134, 2000.

[36] J. A. Kovacs and W. Wriggers. Fast rotational matching, In *Acta Crystallographica*, D58:1282–1286, 2002.

[37] S. Z. Li. Markov random field models in computer vision, *Springer-Verlag*, 2001.

[38] J. S. Lim. Two-Dimensional Signal and Image Processing, *Prentice Hall*, 1990.

[39] S. J. Ludtke, P. B. Baldwin, and W. Chiu. EMAN: Semiautomated Software for High-Resolution Single-Particle Reconstructions, In *Journal of Structural Biology*, 128(1):82–97, 1999.

[40] S. J. Ludtke, J. Jakana, J. L. Song, D. T. Chuang and W. Chiu. A 11.5 åSingle Particle reconstruction of GroEL using EMAN, In *Journal of Molecular Biology*, 314:253–263, 2001.

[41] S. P. Mallick, B. Carragher, C. S. Potter, D. J. Kriegman. ACE: Automated CTF Estimation, In *Ultramicroscopy*, 104(1):8–29, 2005.

[42] D. C. Marinescu and Y. Ji. A computational framework for the 3D structure determination of viruses with unknown symmetry, In *Journal of Parallel and Distributed Computing*, 63:738–758, 2003.

[43] I. A. B. Martin, D. C. Marinescu, R. E. Lynch and T. S. Baker. Identification of spherical virus particles in digitized images of entire electron micrographs, In *Journal of Structural Biology*, 120:146–157, 1997.

[44] J. J. Merelo, A. Prieto, F. Morn, R. Marabini, and J. M. Carazo. Automatic Classification of Biological Particles from Electron-microscopy Images Using Conventional and Genetic-algorithm Optimized Learning Vector Quantization, In *Neural Processing Letters*, 8:55–65, 1998.

[45] J. A. Mindell and N. Grigorieff. Accurate determination of local defocus and specimen tilt in electron microscopy, In *Journal of Structural Biology*, 142:334-347, 2003.

[46] J. Monteil and A. Beghdadi. A new interpretation and improvement of the nonlinear anisotropic diffusion for image enhancement, In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):940–946,1999.

[47] T. K. Moon. The expectation maximization algorithm, In *IEEE Signal Proceedings Magazine*, pp. 47–59, 1996.

[48] F. Mouche, Y. Zho, J. Pulokas, C. S. Potter and B. Carragher. Automated three-dimensional reconstruction of keyhole limpet hemocyanin type 1, In *Journal of Structural Biology*, In Press.

[49] W. V. Nicholson and R. M. Glaeser. Review: automatic particle detection in electron microscopy, In *Journal of Structural Biology*, 133:90–101, 2001.

[50] E. Nogales and N. Grigorieff. Molecular machines: putting the pieces together, In *Journal of Cell Biology*, 152:F1–F10, 2001.

[51] T. Ogura, K. Iwasaki and C. Sato. Topology representing network enables highly accurate classification of protein images taken by cryo electron-microscope without masking, In *Journal of Structural Biology*, 143:185–200, 2003.

[52] T. Ogura and C. Sato. An automatic particle pickup method using a neural network applicable to low-contrast electron micrographs, In *Journal of Structural Biology*, 136:227–238, 2001.

[53] T. Ogura and C. Sato. Automatic particle pickup method using a neural network has high accuracy by applying an initial weight derived from eigen-images: A new reference free method for single particle analysis, In *Journal of Structural Biology*, 145(1-2):63-75, 2004.

[54] E. V. Orlova. Structural analysis of non-crystalline macromolecules: the ribosome, In *Acta Crystallographica*, D56:1253–1258, 2000.

[55] N. Otsu. A threshold selection method from gray level histogram, In *IEEE Transactions on Systems Man and Cybernetics.*, SMC-8:62–66, 1979.

[56] P. Penczek, M. Radermacher and J. Frank. Three dimensional reconstruction of single particles embedded in ice, In *Ultramicroscopy*, 40:33–53, 1992.

[57] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion, In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[58] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition, In *Proceedings of the IEEE*, 77(2):257–286, 1989.

[59] K. Ramani Lata, P. Penczek, and J. Frank, Automatic particle picking from electron micrographs. In *Ultramicroscopy*, 58:381–391, 1995.

[60] A. M. Roseman. Particle finding in electron micrographs using a fast local correlation algorithm, In *Ultramicroscopy*, 94:225–236, 2003.

[61] A. M. Roseman. FindEMa fast, efficient program for automatic selection of particles from electron micrographs, In *Journal of Structural Biology*, 145(1-2):91-99, 2004.

[62] A. M. Roseman and K. Neumann. Objective evaluation of the relative modulation transfer function of densitometers for digitisation of electron micrographs, In *Ultramicroscopy*, 96:207–218, 2003.

[63] M. G. Rossmann and Y. Tao. Cryo-Electron-Microscopy Reconstruction of Partially Symmetric Objects, In *Journal of Structural Biology*, 125:196–208, 1999.

[64] J. Ruprecht and J. Nield. Determining the structure of biological macromolecules by transmission electron microscopy, single particle analysis and 3D reconstruction, In *Progress in Biophysics and Molecular Biology*, 75(3):121–164, 2001.

[65] A. Sali, R. Glaeser,T. Earnest and W. Baumeister. From words to literature in structural proteomics, In *Nature*, 422:216–225, 2003.

[66] B. Sander, M. M. Golas and H. Stark. Corrim-based alignment for improved speed in single-particle image processing, In *Jorunal of Structural Biology*, 143:219–228, 2003.

[67] B. Sander, M. M. Golas and H. Stark. Automatic CTF correction for single particles based upon multivariate statistical analysis of individual power spectra, In *Jorunal of Structural Biology*, 142:392–401, 2003.

[68] T. R. Shaikh, R. Hegerl, and J. Frank. An approach to examining model dependence in EM reconstructions using cross-validation, In *Journal of Structural Biology*, 142:301–310, 2003.

[69] C. E. Shannon. Communication in the presence of noise, In *Proceedings of IRE*, 37:10–22, 1949.

[70] J. M. Short. SLEUTH - A fast computer program for automatically detecting particles in electron microscope images, In *Journal of Structural Biology*, 145:100–110, 2004.

[71] F. J. Sigworth. Classical detection theory and the cryo-EM particle selection problem, In *Journal of Structural Biology*, 145:111–122, 2004.

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

[73] V. Singh, Y. Ji, and D. C. Marinescu. A Parallel Algorithm for Automatic Particle Identification in Electron Micrographs, In *VECPAR 2004, LNCS*, 3402:354–367, 2005.

[74] P. Soille. Morphological Image Analysis: Principles and Applications, In *Springer-Verlag*, 1999.

[75] Y. Song and A. Zhang. Monotonic Trees, *10th International Conference on Discrete Geometry for Computer Imagery*, 114–123, 2002.

[76] C. O. S. Sorzano, R. Marabini, J. Velzquez-Muriel, J. R. Bilbao-Castro, S. H. W. Scheres, J. M. Carazo and A. Pascual-Montano XMIPP: a new generation of an open-source image processing package for electron microscopy, *Journal of Structural Biology*, 148(2):194–204, 2004.

[77] C. Suloway, J. Pulokas, D. Fellmann, A. Cheng, F. Guerra, J. Quispe, S. Stagg, C. S. Potter and B. Carragher Automated molecular microscopy: The new Leginon system, *Journal of Structural Biology*, in press (Available online at http://dx.doi.org/10.1016/j.jsb.2005.03.010 on 10 May 2005), 2005.

[78] P. A. Thuman-Commike, W. Chiu. Reconstruction principles of icosahedral virus structure determination using electron cryomicroscopy, In *Micron*, 31:687–711, 2000.

[79] Y. Cong, J. A. Kovacs and W. Wriggers. 2D fast rotational matching for image processing of biophysical data, In *Journal of Structural Biology*, In Press.

[80] M. van Heel. Detection of objects in quantum-noise-limited images, In *Ultramicroscopy*, 7(4):331–341, 1982.

[81] M. van Heel. Angular reconstitution: a posterori assignment of projection directions for 3D reconstruction, In *Ultramicroscopy*, 21:111–124, 1987.

[82] M. van Heel, B. Gowen, R. Matadeen, E. V. Orlova, R. Finn, T. Pape, D. Cohen, H. Stark, R. Schmidt, M. Schatz, and A. Patwardhan. Single-particle electron cryo-microscopy: towards atomic resolution, In *Quarterly Reviews of Biophysics*, 33(4):307–369, 2000.

[83] M. van Heel and G. Harauz. Resolution criterion for three dimensional reconstruction, In *Optik*, 73:119–122, 1986.

[84] M. van Heel, G. Harauz, E. V. Orlova, R. Schmidt, and M. Schatz. A New Generation of the IMAGIC Image Processing System, In *Journal of Structural Biology*, 116(1):17-24, 1996.

[85] N. Volkmann and D. Hanein. Quantitative fitting of atomic models into observed densities derived by electron microscopy, In *Journal of Structural Biology*, 125:176–184, 1999.

[86] J. Weickert. *Anisotropic Diffusion in Image Processing*, Teubner-Verlag, 1998.

[87] Y. Zhang, S. Smith, and M. Brady. Segmentation of brain MRI images through a hidden Markov random field model and the expectation-maximization algorithm, In *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

[88] Z. Yin, Y. Zheng and P. C. Doerschuk. An ab initio algorithm for low resolution 3D reconstructions from cryoelectron microscopy images, In *Journal of Structural Biology*, 133:132–142, 2001.

[89] Y. Yitzhaky, and E. Peli. A Method for Objective Edge Detection Evaluation and Detector Parameter Selection, In *IEEE Pattern Analysis and Machine Intelligence*, 25(8):1027–1033, 2003.

[90] Z. Yu, and C. Bajaj. Detecting circular and rectangular particles based on geometric feature detection in electron micrographs , In *Journal of Structural Biology*, 145:168–180, 2004.

[91] F. Zemlin. Desired features of a cryoelectron microscope for the electron crystallography of biological material, In *Ultramicroscopy*, 46:1–4, 1992.

[92] Y. Zhu, et. al. Automatic particle selection: results of a comparative study , In *Journal of Structural Biology*, 145:3–14, 2004.

[93] Y. Zhu, B. Carragher, D. Kriegman, F. Mouche, and C. S. Potter. Automatic particle detection through efficient Hough transforms, In *IEEE Transactions on Medical Imaging*, 22(9):1053–1062, 2003.

[94] Y. Zhu, B. Carragher, D. Kriegman, R. A. Milligan, and C. S. Potter. Automated identification of filaments in cryoelectron microscopy images, In *Journal of Structural Biology*, 135(3):302–312, 2001.

[95] Y. Zhu, P. Penczek, R. Schroder and J. Frank. Three-Dimensional Reconstruction with Contrast Transfer Function Correction from Energy-Filtered Cryoelectron Micrographs: Procedure and Application to the 70SEscherichia coli Ribosome, In *Journal of Structural Biology*, 118(3):197–219, 1997.