Automatic feature extraction and statistical shape model of the AIDS virus spike

Ajay Gopinath* and Alan C. Bovik

Abstract—**We introduce a method to automatically extract spike features of the AIDS virus imaged through an electron microscope. The AIDS virus spike is the primary target of drug design as it is directly involved in infecting host cells. Our method detects the location of these spikes and extracts a subvolume enclosing the spike. We have achieved a sensitivity of 80% for our best operating range. The extracted spikes are further aligned and combined to build a 4D statistical shape model, where each voxel in the shape model is assigned a probability density function. Our method is the first fully automated technique that can extract sub-volumes of the AIDS virus spike and be used to build a statistical model without the need for any user supervision. We envision that this new tool will significantly enhance the overall process of shape analysis of the AIDS virus spike imaged through the electron microscope. Accurate models of the virus spike will help in the development of better drug design strategies.**

Index Terms—**Feature Extraction, Statistical Shape Analysis, Electron Microscopy, AIDS Virus, Spike gp120**

I. Introduction

The AIDS virion is roughly spherical in shape, with an inner capsid region that encloses its genome and an outer proteinaceous envelope on which several protruding entities called *spikes* are distributed. Each spike is roughly mushroom shaped with a tapering stem that is attached to the virus envelope. The head of the mushroom shaped structure has a trimeric protein known as gp120, each of whose monomeric subunits is arranged symmetrically around an axis passing through the center of the spike. A cylindrically shaped protein known as gp41 connects with the proteinaceous envelope. The virus particle is typically $120nm$ in diameter while the height of the spike is around $120\AA$ with a maximum width of about $150\AA$, tapering to $35\AA$ at the junction of the envelope [1].

The spike is the primary target for drug design as it allows the virus to infect the immune cells by binding and fusing with them. The precise structure and various possible states of the virus spike is of high importance for biochemists who design drugs that can neutralize the AIDS virus. Shape complementarity between the drug and the virus spike is one of the critical aspects of drug design. Currently, biochemists identify spikes and segment them through manual supervision or by semi-automated methods where a user provides the initial locations or inputs to a segmentation algorithm that extracts spike features within a user defined subvolume [1] [2]. Liu et al. [1], Zhu, et al. [3] and others have extracted several individual spikes using manual processes and performed alignment and averaging to create a single averaged spike model [1]. Spike extraction methods involving user supervision can be tedious and time consuming. Our objective is to use image processing and computer vision to fully automate the process of detecting and extracting spikes without the need for any user supervision. We demonstrate the efficacy of our method by also producing a statistical shape model of the spike instead of a single average shape model as reported in the literature [1], [3]. This fully automated process could significantly enhance the overall spike analysis pipeline, providing biochemists involved in drug design the ability to process virus data in much larger numbers leading to more accurate structure elucidation. To the best of our knowledge our algorithm is the first method that directly addresses the spike detection and statistical model generation in a fully automated framework.

We use imaging data from a Transmission Electron Microscopy (TEM), which is the preferred tool for structural biologists to visualize three-dimensional structures of molecular and cellular complexes *in-situ*. Electron tomography (ET) involves acquiring planar TEM images of the biological sample from different projection angles (tilt series) and reconstructing a 3D volumetric image (or map) from these projections using the principles of tomography. This is currently the *only* approach that allows one to reconstruct the 3D structure of individual biological complexes in their native state. TEM images suffer from a limited contrast to noise ratio. A second major challenge is that the angle of rotation cannot exceed $\pm 70^\circ$ (with respect to the horizontal plane) because the beam's path-length through the sample and its supporting structure becomes inappropriate to form projections. Hence projections are available only for a limited number of tilt angles. In Fourier space, the missing tilt slices at higher angles appear as a wedge and this is also known as the missing wedge problem. This results in severe blurring of the biological structure being imaged, making the virus isolation, identification and modeling problems much more difficult.

A. Statistical Shape Models

A single template shape is not sufficient for most biological structures due to their high variability. A statistical model aims to include common variations of the structure in the model. The most common and simplest method to represent shapes is a set of points that are distributed across the structure's surface. These points are commonly referred to as landmarks, though they need not be located at salient feature points as per the common definition for anatomic landmarks [4]. Landmarks have been used to build statistical shapes of biological

Asterisk indicates corresponding author.

The authors are with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712, USA. email: ajay.gopinath@utexas.edu, bovik@ece.utexas.edu

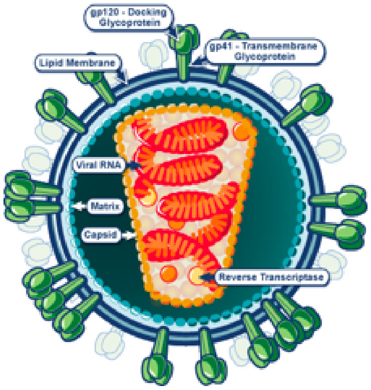

gp120 - Docking Glycoprotein
gp41 - Transmembrane Glycoprotein
Lipid Membrane
Viral RNA
Matrix
Capsid
Reverse Transcriptase

Fig. 1: A schematic of the HIV virus shown here. The virus spikes are the gp120 and the gp41 regions that protrude out of the virus envelope. (source: www.niaid.nih.gov)

structures by Bookstein [5] and others. Medial axis models or skeletons have also been used to describe biological shapes. The structure is represented by centerlines and corresponding radii. Pizer et al. [6] introduced a medial model with a coarse-to-fine representation that uses a collection of points on centerlines and vectors towards the boundary. A non-uniform rational B-Splines (NURBS) method was used by Tsagaan et al. [7] to model a variety of objects, including the kidney that possess intricate features. Methods that use landmarks need to ensure that they are all located on corresponding positions across all the training samples. Obtaining the correspondence of landmarks across several 3D volumes is not trivial. Typical methods to construct a statistical shape involve extracting a mean shape and modes of variation from training samples. The first step is to align the shapes; there are several methods for aligning both rigid and non-rigid objects. The aligned data can be compactly represented using methods such as principal component analysis (PCA) that represent shapes by a linear combination of modes. In general, PCA results in global modes which influence all variables simultaneously, hence varying one model will affect the entire shape [4].

Other methods include those by Cootes and Taylor [8] that use finite element methods to calculate vibrational modes for the training data and that are used to create a model that can represent all shape instances. To increase the model flexibility, some approaches split the statistical shape into different parts, where each part varies independently. Zhao et al. [9] create a multi-partite model by using mesh partitioning, where each part of the mesh is modeled separately. Rueckert et al. [10] employ statistical deformation models (SDM) to construct anatomical models of the brain. This method is closely related to the developing field of Computational Anatomy (CA) method promoted by Grenander and Miller [11]. Computational Anatomy involves generating models from a set of anatomical images [12] [13] [11]. The idea in CA is to carry out statistical analysis directly on deformation fields that are obtained by performing non-rigid registration, without the need for segmentation and correspondence estimates. Also, instead of performing analysis on the deformation field directly, the statistical analysis is performed on the control points of the deformation fields. The advantage of these control points is in providing compact representation of the deformation field. Methods described by Rohlfing et al. [14] use repeated application of a non-rigid registration method based on B-splines to generate an average model.

To build a statistical shape model of the AIDS virus spike, we use spike sub-volume that are extracted and aligned automatically. We combine intensity information from all the individual detected spikes. The resulting statistical model of the spike is 4D, where the fourth dimension is a probability density function assigned for that voxel. The density function is constructed at each voxel based on the samples from all the detected spikes.

II. Method

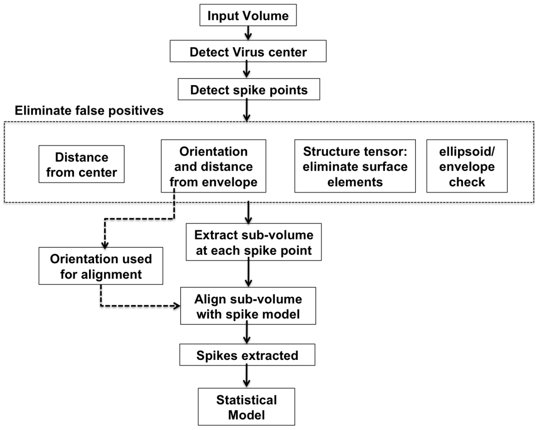
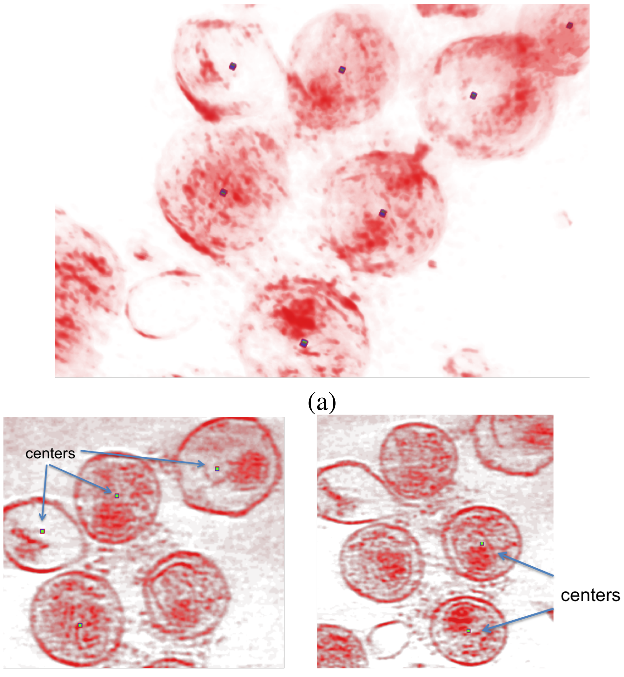

Input Volume
Detect Virus center
Detect spike points
Eliminate false positives
Distance from center
Orientation and distance from envelope
Structure tensor: eliminate surface elements
ellipsoid/ envelope check
Extract sub-volume at each spike point
Orientation used for alignment
Align sub-volume with spike model
Spikes extracted
Statistical Model

Fig. 2: Algorithm flow of the spike detection and model generation method.

Volumetric images are generated from the tomographic reconstruction of single axis tilt series images taken from the range $\pm 69^\circ$ from a TEM. The images are of the Simian Immunodeficiency Virus (HIV-like retrovirus that causes AIDS in monkeys) [15]. We used a Maximum Likelihood reconstruction scheme to perform the tomographic reconstruction [16]. The reconstructed volume is of size $512 \times 512 \times 512$ and contains about 9 virus particles, approximately 70 voxels in diameter. The approximate bounding box of the spike head (gp120) is $10 \times 10 \times 10$ voxels. The overall bounding box of the entire spike including the head of the spike (gp120) the tapering stem (gp41) and the adjoining virus envelope is $10 \times 10 \times 14$ voxels. The overall flow of the algorithm is shown in Fig. 2. It begins by detecting the center of each putative virus particle then extracting a sub-volume that contains the virus particle. Candidate points that may lie on the spike are detected, and false positives are eliminated based on a number of specific physical criteria. Sub-volumes of the spike are extracted at each point, then are aligned and combined to create a statistical model.

A. Spike Detection

1) Detect virus center: The first step of spike detection is to identify the centers of the virus particles. Since these have a roughly spherical outer envelope, we use a template matching technique to detect the spherical envelope region. We created 64 ellipsoid templates with radii varying from 33 to 36 voxels along the x, y and z dimensions. These ellipsoids were hollow with an outer shell of thickness 2 voxels corresponding to the width of the virus envelope. As a pre-processing step, the input volume containing the virus particles is thresholded with a very low value that eliminates low intensity background regions. The normalized cross correlation (NCC) is then calculated in the frequency domain for each of the ellipsoid templates with the input volume containing the virus particles. This results in 64 NCC volumes. The local maxima, in a $4 \times 4 \times 4$ region, over all of the 64 NCC volumes are selected as centers. Centers that lie near the volume edge are eliminated as false positives. The centers of all the virus particles were successfully captured using this method. A sub-volume of size $100 \times 100 \times 100$ containing the entire virus particle centered on the detected virus center is then extracted from each data volume.



(a)

(b)

(c)

Fig. 3: Input volume with detected virus centers (a) 3D volume rendering of the virus particles with the detected virus centers depicted as dots inside. (b) and (c) 2D slices: The centers of 4 virus particles occur on the same 2D slice and are shown in (b). The center of the fifth virus occurs on another 2D slice. The centers of 2 virus particles on the same 2D slice are shown in (c). The centers of the remaining virus particles are on other slice planes.

2) Detect spike-points: In this step we detected candidate points that lie on a virus spike in each virus sub-volume that was extracted previously. The head of the spike gp120 region is a blobby shaped structure. We used a difference of Gaussian (DoG) operator to identify the blobby regions by selecting the local maxima of the DoG responses as candidate points. These are points that lie on blobby structures, including spikes. We refer to these candidate points as *spike-points*. The DoG is a close approximate of the second derivative of a Gaussian (Laplacian of Gaussian). Evaluating the DoG involves subtracting two different scales of the sub-volume region of the virus particle. We created a scale space of 5 volumes by convolving the original with a Gaussian kernel at $\sigma = [0.707, 1.41, 2.12, 2.828, 3.355]$. Four DoG volumes are generated by subtracting two consecutive scales, i.e. volumes at $\sigma = 0.707,\ 1.41$ are used to generate one DoG volume and $\sigma = 1.41,\ 2.12$ are used to generate another and so on (Fig. 4).

a) Local Maxima of DoG:: Local maxima are located for each DoG volume and its immediate neighbors. The local maxima check is performed for each voxel's 26 neighbors in the current DoG volume and the DoG volume above and below it (see Fig. 4). A voxel in DoG-2 is compared against its 26 neighbors and the 27 voxels in the corresponding location of DoG-1 and DoG-3. Similarly DoG-3 voxels are compared with DoG-2 and DoG-4. The resulting local maxima points are referred to as spike-points and are candidate points on the spikes. Typically, we obtain about $1,000$ spike-points for a single virus particle. The next steps attempt to eliminate the false positives and preserve only those points that lie on a spike.

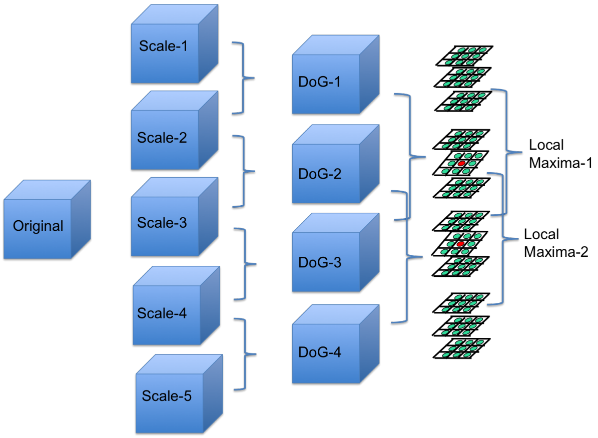


Fig. 4: Detecting spike-points: The subvolume containing the virus particle is scaled by convolution with a Gaussian kernel at different sigmas. Difference of Gaussian (DoG) volumes are computed by subtracting neighboring scaled volumes. Local maxima of each DoG volume including the local neighborhood of the adjoining DoG volumes are identified. These local maxima are points which lie on blobby regions of the original volume and are called spike-points.

3) False positive reduction: We used an array of stages to eliminate false positives from the detected spike-points. A soft threshold approach was used, by defining a confidence range $[0, 1]$, where 1 indicates high confidence. Confidences are assigned to each spike point at every false positive reduction stage. The decision on whether a point is a false positive is made at the end by combining the confidence values from all the stages.

a) Distance from virus center: Based on the current literature about the structural characteristics of the AIDS virus and the typical virus radii seen in our data, we observed that spikes on the virus envelope occur at least 30 voxels from the approximate virus center. Spike-points that lie less than 30

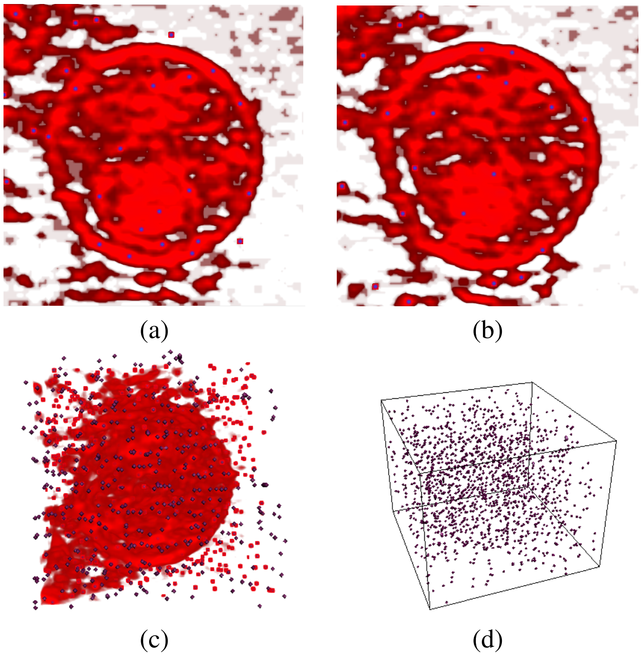

(a)
(b)
(c)
(d)

Fig. 5: Spike-points before FP removal: (a) and (b) are 2D slices of a virus particle with spike-points shown in blue. (c) is a 3D volume rendering of the virus particle with spike-points. (d) shows the spike-points in 3D in a sub-volume containing a virus-particle (not shown).

voxels from the virus center are most likely false positives. We assigned a soft-threshold value of 1 for all spike-points further than 30 voxels from the center. Those that are below 30 are assigned a confidence value of $1 - \frac{30-distance}{30}$. Points that are very close to the center, at about 20 voxels, are rejected.

b) Distance from envelope and orientation: Given a spike-point, we calculated the orientation of the spike-point with respect to the virus envelope and estimated its approximate distance from the virus envelope. Spikes typically protrude radially from the virus envelope (Fig. 7). As explained in the description of the virus center detection (Section II-A1) the algorithm finds the ellipsoid that best correlates with each virus particle. Given the ellipsoid parameters, the normal from the surface of the ellipsoid to the spike point is calculated. Let $\vec{p}$ be the spike-point and S the ellipsoid surface. Then the vector $\vec{x}$, such that the spike axis $\vec{px}$ is normal to the ellipsoid surface S at $\vec{x}$ (Fig. 6) satisfies

$$(\vec{p} - \vec{x})\vec{x}' = 0, \qquad (1)$$

where $\vec{x}'$ is the tangent at $\vec{x}$ and $\vec{p} - \vec{x}$ is the normal. The vector $\vec{x}$ may be parametrized in spherical coordinates for computational efficiency [17]:

$$x(\vec{\theta}, \phi) = r \begin{bmatrix} \alpha cos(\phi) sin(\theta) \\ \beta cos(\phi) sin(\theta) \\ \gamma sin(\phi) \end{bmatrix} \qquad (2)$$

where α, β, γ and r represent the parameters of the ellipsoid.

Proceeding directly, solve for θ and ϕ in the set of equations

$$f := \begin{cases} (\vec{p} - \vec{x})\frac{\partial \vec{x}}{\partial \theta} = 0 \\ (\vec{p} - \vec{x})\frac{\partial \vec{x}}{\partial \phi} = 0, \end{cases} \qquad (3)$$

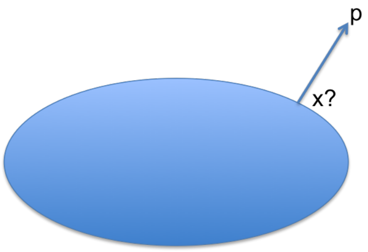


Fig. 6: Spike orientation axis: Given a spike-point $\vec{p}$, the objective is to find $\vec{x}$ such that the spike orientation axis $\vec{px}$ is normal to the surface of the ellipsoid.

by using Newton's method:

$$\begin{bmatrix} \Delta\theta \\ \Delta\phi \end{bmatrix} = -J^{-1} f, \qquad (4)$$

where J is the Jacobian of f.

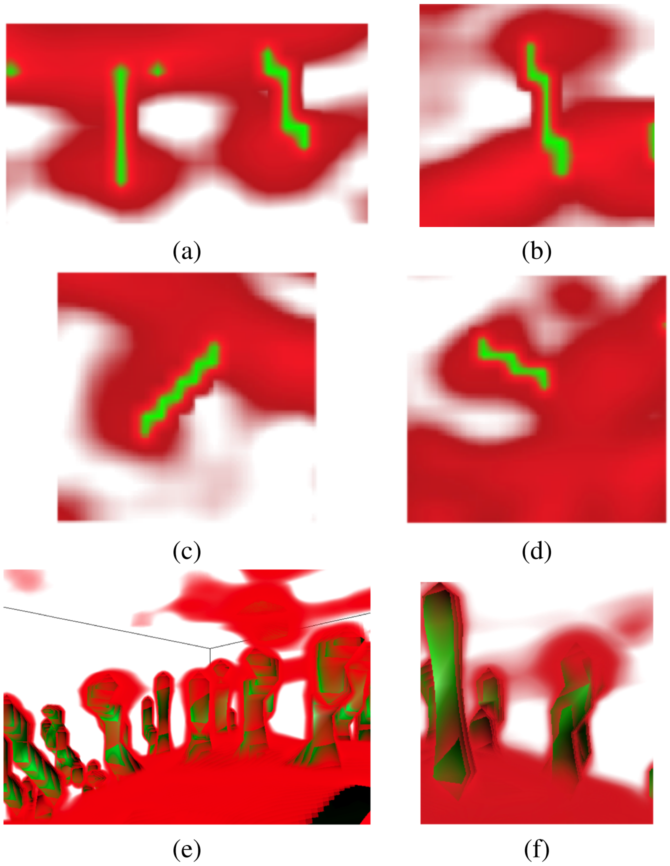
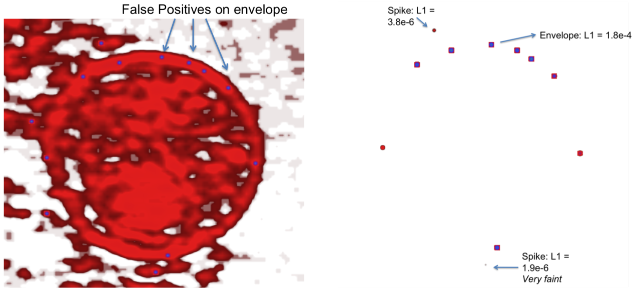

(a)
(b)
(c)
(d)
(e)
(f)

Fig. 7: Spike orientation axis: (a, b, c, d) are 2D slices with the detected spike orientation axis shown in green. (e, f) are 3D volume renderings of the spike region with the detected axis rendered in green.

The iterations begin with an initial estimate $\theta = tan^{-1}(\frac{\alpha p_x}{\beta p_y})$ and $\phi = tan^{-1}(\frac{p_z}{c\sqrt{(\frac{x}{a})^2+(\frac{y}{b})^2}})$ and are updated with $\Delta\theta$ and $\Delta\phi$. Convergence is obtained in about 3 iterations. The θ and ϕ parameters found using Newton's method was used to estimate x as given in (2). Using this formulation,



(a) (b)

Fig. 8: Structure tensor for false positive elimination: (a) Shows false positives on the virus envelope that are eliminated by evaluating the structure tensor. (b) Points on the envelope have a large L_1 value compared to the ones on the spike, which are blobby. False positives are eliminated based on this.

a likely orientation axis of each spike $\vec{px}$ and its length $|\vec{px}|$ is computed. Note that the actual orientation of the spikes may differ slightly from the calculated orientations since the virus envelope is approximated by an ellipsoid and the spike orientation may not be exactly normal to the envelope surface. The orientation axis estimated using this approach is a good initial estimate for the alignment and spike extraction performed in Section II-A4. The length $|\vec{px}|$ is used to eliminate spike points which lie too far away from the surface. A confidence value of 1 is given to spike-points that are less than 10 voxels in length. If the distance is greater than 10 then a confidence value of $1 - \frac{|length-10|}{10}$ is assigned. Spike points that are greater than the length threshold by a tolerance (here 25%) are eliminated. A fast check to eliminate spike-points that lie inside the ellipsoid and hence the virus is performed. The distance of the spike point $\vec{p}$ from the virus center is compared with the distance of the ellipsoid point $\vec{x}$ from the center. If the distance of $\vec{p}$ to the center is smaller than $\vec{x}$ to the center, then it implies that the point lies inside the ellipsoid and is eliminated.

c) Fitting ellipsoid on envelope: We use the ellipsoid template to detect false positive spike points that lie on the virus envelope. For each virus particle, the ellipsoid that gave the highest NCC value for its center (Section II-A1) is chosen as a template for the virus. Affine registration is performed between the ellipsoid and the corresponding virus to align them. Spike points that lie on the registered ellipsoid or inside it are tagged as false positives. A false-positive membership value is then assigned based on the fraction of neighbors of the spike-point that are inside the ellipsoid.

d) Structure tensor: Spike-points that lie on the virus envelope were a large source of false positives. Structure tensors are used to detect points that lie on the envelope. At each spike-point, the structure tensor is calculated and points that were on a surface-like structure are eliminated while preserving those that lie on blobby regions. To make the structure tensor calculation more robust, local region growing is performed in a $5 \times 5 \times 5$ region around the spike-point, thereby enabling the computation of partial derivatives on the points extracted.

$$[I_x, I_y, I_z] \;=\; \nabla(\vec{p}) \tag{5}$$

$$StructureTensor \;=\; \begin{bmatrix} I_x^2 & I_xI_y & I_xI_z \\ I_xI_y & I_y^2 & I_yI_z \\ I_xI_z & I_yI_z & I_z^2 \end{bmatrix} \tag{6}$$

$$[L_1, L_2, L_3] = Eigen(StructureTensor) \tag{7}$$

Here I_x, I_y, I_z are the partial derivatives at spike point $\vec{p}$ and L_1, L_2, L_3 are the eigen values of the structure tensor. Spike points that lie on the envelope of the virus have a surface-like neighborhood with a “surfel” (Surface-Element) characteristic, where:

$$L_1 \gg L_2 \approx L_3 \tag{8}$$

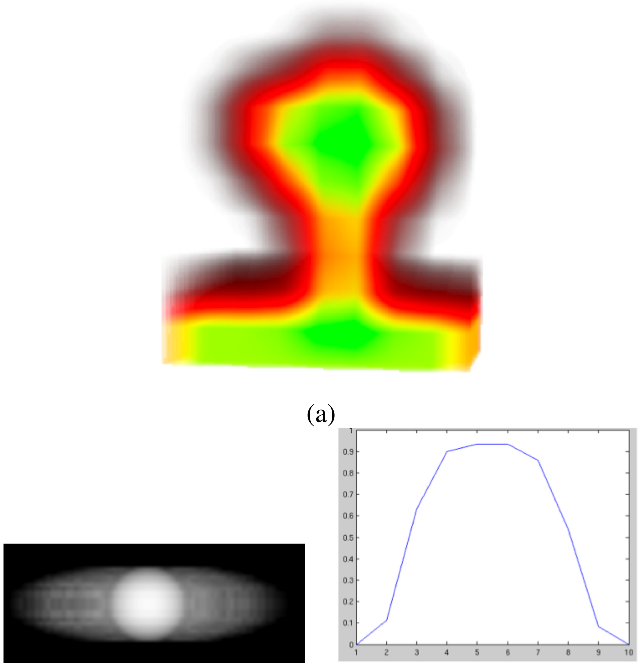
L_1 for points that were on the envelope $\geq 10^{-4}$ while those that were on spikes are observed to be $\leq 10^{-6}$. L_2, L_3 are much lower, on the order of 10^{-8}. Points with $L_1 \leq 2 \times 10^{-4}$ (threshold) are assigned a confidence level of 1 while those that are greater than the threshold are assigned a value of $1 - \frac{|L_1 - threshold|}{threshold}$. Points with a value beyond the threshold by a tolerance of 25% are eliminated as false positives. Points on spikes, which are blobby and have much lower L_1, are preserved.

At this stage the algorithm has eliminated a large percentage of false positives, yielding about 150 spike points and a set of membership values for each point based on the false positive elimination stages. The membership values from each stage are averaged and assigned as the final membership value for each spike point. The final membership values of most spikes are close to 1 with about 20% of the points having membership values ranging from $0.75\ to\ 1$. We plotted a Free-response Receiver Operating Characteristic (FROC) to estimate the optimum choice of threshold of the final membership value to eliminate false positives (Section III-A).

4) Extracting spikes: Our next step was to extract spikes at the detected spike points and align them to build a statistical model.

a) Phantom spike: We created a phantom structure shaped like a virus spike to aid in extracting the virus spikes. A cylindrical structure with a sphere placed on top is created to replicate the shape of a virus spike. This model is blurred with a Gaussian. The blur parameters are estimated by simulating projections of a sphere between $\pm 69^\circ$ in steps of 3° and reconstructing it through back projection. The resulting sphere with blur due to limited angle tomographic effects is used to estimate the σ value used for blurring the spike model (Fig. 9). For each spike point, the extracted phantom spike is aligned along the spike axis predicted in Section II-A3b.

b) Extract subvolume and its orientation: At each spike point that has filtered through the false positive removal process, a $10 \times 10 \times 14$ sub-volume, the observed size of a typical spike (see Section II), is extracted. The sub-volume's intensity range is normalized to lie in the interval $[0, 1]$ and processed using thresholding and connected component analysis. The choice of threshold varies from $0.45 - 0.2$ where the appropriate threshold is selected based on the FROC analysis (see Section III-A). At high threshold values, spike points



(a)

(b)

(c)

Fig. 9: Phantom model used for extracting a spike: (a) Blurred spike model built using ellipsoid on top of a cylinder and an envelope region at the base. (b) 2D slice of reconstructed (backprojection) sphere phantom, used to estimate blurr parameters. A Gaussian blur with the estimated parameters is applied to the ellipsoid-cylinder model. (c) 1D intensity profile through the center of the 2D slice shown in (b).

that lie on blobby regions with very weak intensity regions can break up into several small connected components. These spike-points are eliminated as false positives as a reliable spike region cannot be extracted. While spike points that lie on spikes with good contrast and that are distinct from the background produce a large connected component that includes the spike-point, these are preserved. This sub-volume is next compared with the phantom spike model in order to recover its orientation.

An affine registration between the sub-volume region and the phantom spike oriented along the predicted spike axis is performed. This is to recover only small rotation and translation shifts between the predicted axis and the actual axis of the spike. Large rotations that would invert the orientation of the spike are thereby prevented. With this, the orientation of the spike present in the sub-volume was estimated.

The spike point is shifted by ± 1 along each dimension and the spike sub-volume extraction process described above is repeated on each shifted spike-point. The sub-volume that delivers the best similarity match with the phantom model is selected. We use a wavelet based structure similarity index developed by Sampat et al. [18] to compute the similarity between the sub-volume and the phantom model. Structure similarity measures provide more robust matching than metrics such as mean square error.

This procedure described above was performed for all the spike points and we extracted the sub-volume containing the spike and also determined its orientation (Fig. 10). With the orientation established, it is now possible to combine all the extracted spikes in the next step of building a statistical shape model.

B. Building a Statistical Shape Model

The spikes extracted from the previous step (Section II-A4b) are all aligned in a common coordinate frame. The spikes in the AIDS virus exhibit three-fold symmetry about the central axis. The pose of each spike is recovered by rotationally aligning it with the spike model used by Liu et al. [1]. This is done only to recover the pose or rotation of the spike along the axis of the spike. At this stage all the spikes are fully aligned with each other.

To build the statistical model, the information from all the aligned individual spikes is combined to form a probability density function at each voxel. The statistical model is in 4D space where the fourth dimension is the probability density function. The density function was constructed at each voxel based on the intensity values of the detected spikes at that voxel. We used a kernel smoothing density estimate to construct the density function at each voxel.

C. Spike Membership and Model Refinement

Each detected spike's voxel has a membership value in the overall statistical model. We built a membership volume for each spike, where each corresponding voxel is assigned an intensity value equal to the membership of that voxel in the statistical model (Fig. 10). Noisy regions of the spike have low membership values. We reported an average membership number for each membership volume. This gives an overall membership value of a spike in the statistical model. We noticed that spikes that are blurry and noisy have lower average membership values while those that are more distinct have higher average membership values.

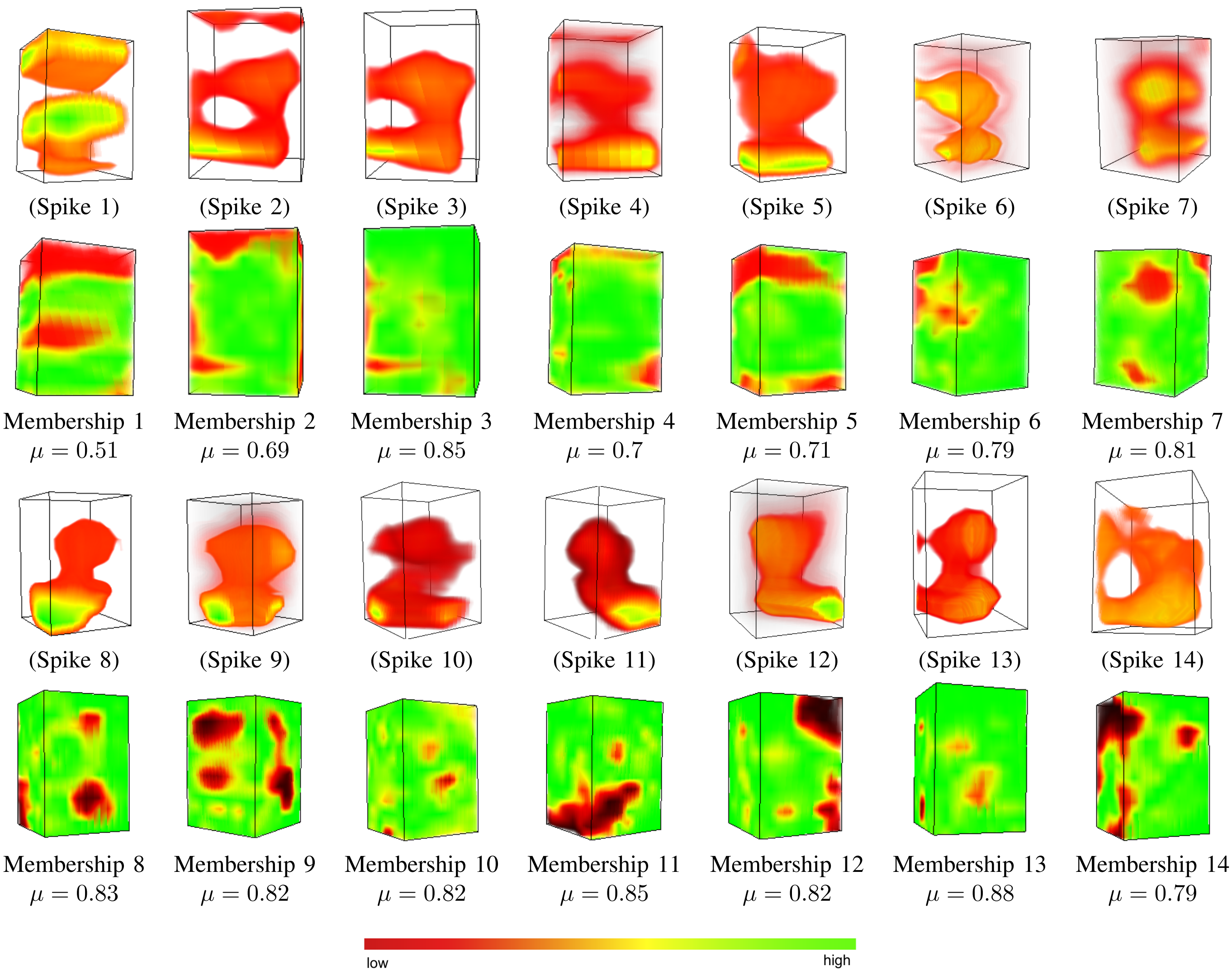
III. Results and discussion

A. FROC Analysis

To analyze our spike detection algorithm, we performed a Free-response Receiver Operator Characteristic (FROC) study [19]. A FROC curve is a plot of sensitivity vs the number of false positive detections per virus. Sensitivity is defined as the fraction of spikes detected.

$$Sensitivity = \frac{Number\ of\ True\ Positives}{Total\ number\ of\ spikes} \qquad (9)$$

For this study, ground truth for 96 spikes in four different virus particles were marked. A knowledgeable user marked a point on the head of the spike's approximate center. The detection and extraction algorithm was run and the resulting location of spike-points compared against the ground truth. A spike was considered detected if a spike-point was placed by the algorithm within a radius of 7 voxels from the marked ground truth point (acceptance radius). This was chosen because the head of the spike gp120 region is about $10 \times 10 \times 10$ voxels



(Spike 1) (Spike 2) (Spike 3) (Spike 4) (Spike 5) (Spike 6) (Spike 7)

Membership 1 $\mu = 0.51$ Membership 2 $\mu = 0.69$ Membership 3 $\mu = 0.85$ Membership 4 $\mu = 0.7$ Membership 5 $\mu = 0.71$ Membership 6 $\mu = 0.79$ Membership 7 $\mu = 0.81$

(Spike 8) (Spike 9) (Spike 10) (Spike 11) (Spike 12) (Spike 13) (Spike 14)

Membership 8 $\mu = 0.83$ Membership 9 $\mu = 0.82$ Membership 10 $\mu = 0.82$ Membership 11 $\mu = 0.85$ Membership 12 $\mu = 0.82$ Membership 13 $\mu = 0.88$ Membership 14 $\mu = 0.79$

Fig. 10: Automatically extracted spikes and their membership volumes according to the statistical model. The statistical model has a probability distribution associated with each voxel. We computed a membership volume where each voxel in the membership volume corresponds to the membership value of the spike's voxel in the statistical model. The mean membership value for each membership volume is also shown. Noisy regions of the spikes tend to have lower membership values at those regions.

(see spike description in Section II), which corresponds to a midradius (center to edge) of about 7 voxels. Multiple spike-point detection within this radius was considered as a single detection. The detection algorithm's final membership value and the threshold value used in the extraction of spikes were varied and the resulting sensitivities was plotted against the false positive rate.

The best sensitivity was at 0.81 where 77 out of 96 spikes were detected with 9 false positives (FP) per virus. At 0.75 sensitivity, 72 out of 96 spikes were detected and only 7 FP per virus. This seems to be the best operating range for the detection algorithm. Below this point, the sensitivity drops sharply. When we completely relaxed the false positive elimination stages, 99% of the spikes were detected. But this also yielded a large 56 false positives per virus. At 93% sensitivity 26 false positives were found per virus. Our best operating range is at a sensitivity of about 75 to 80% with 7 to 9 false positives per virus.

B. *Statistical Shape Model*

The statistical model is in 4D data where the fourth dimension represents the probability density function associated with each voxel. Visualizing and displaying such data in its entirety is challenging. Fig. 12 shows a 3D volume rendering and a central 2D image along the XZ plane of the mean statistical model. Each voxel in the mean statistical model has an intensity value that is the mean value of the detected spikes at that voxel. We have displayed 2D profiles of the average spike model and plotted the density function on a set of voxels. Figs. 13(a) and (b) show a profile image of the average statistical model along the XZ plane and plots of the density function associated with voxels that lie on the center line shown in the image. Fig. 13(c) is an image along the XY plane of the average statistical model and the associated density function plots are shown. Voxels that lie near the center of the image have a bigger spread in their density function and larger mean values than those that lie near the boundary. A

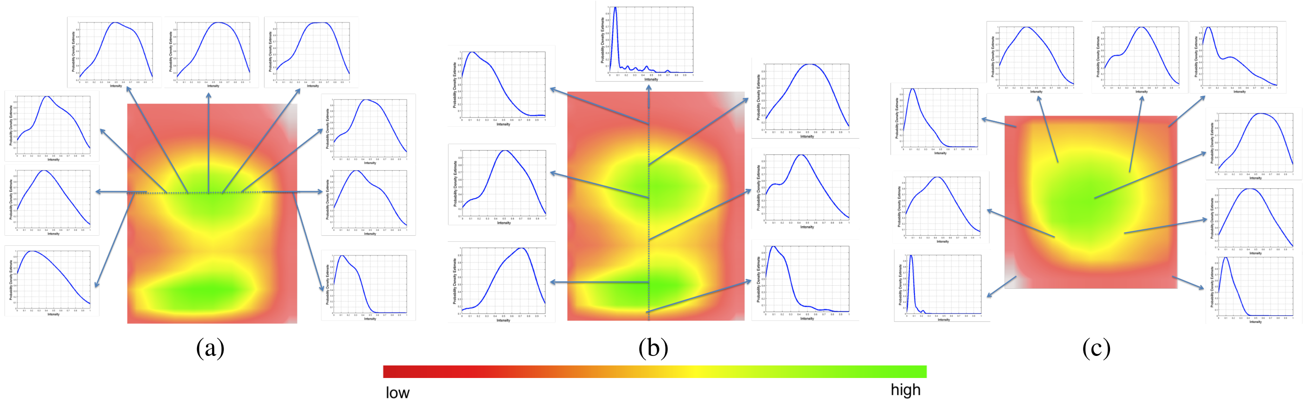


Fig. 13: Statistical Shape Model: (a) and (b) show a profile image of the mean statistical model along the XZ plane and plots of the density function associated with voxels that lie on the center line shown in the image. (c) XY plane of the mean statistical model and the associated density function plots are shown. Voxels that are at the center have higher mean values (green), while those near the edges have lower mean values (red).

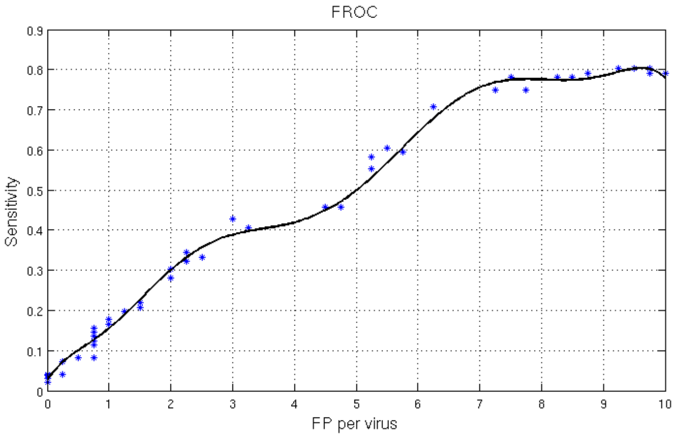


Fig. 11: Free response Receiver Operating Characteristic (FROC) curve plots the sensitivity vs the number of false positive detections per virus. We measured 80% sensitivity (80% of all spikes were detected) with about 9 false positives (FP) per virus. Our best operating range is at 7 FP with a sensitivity around 75%, beyond which the sensitivity drops.

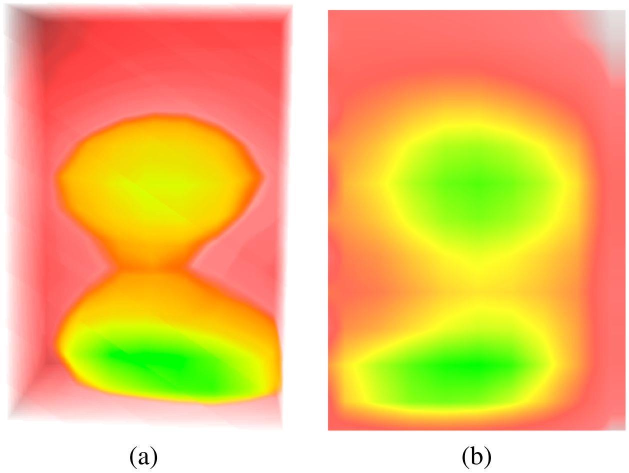


Fig. 12: (a) Volume rendering (VR) of the mean statistical shape model of the AIDS virus spike. (b) is a 2D slice of the mean statistical shape model.

small spread for voxels further from the center along with a low mean value implies a high confidence (low uncertainty) bound for the size and shape of the spike. As seen in Fig. 13(b), there is a higher uncertainty associated with the stem region (gp41) that connects the head of the spike (gp120) to the envelope. Whereas, the head of the spike region (gp120) and the envelope have greater confidences and higher mean values.

C. Applications

The statistical model can be used for various computational biology applications. Some important applications are discussed here:

1) Fitting: Using X-ray crystallography studies, biologists can determine the atomic structures of the subcomponents of biological complexes. A drawback is that the biological complex is not in its native state since these atomic structure models are obtained using crystal lattice growing techniques. A common practice is to perform fitting [20], where the atomic model is aligned with a coarser model obtained from EM imaging, in which the biological complex is in its native state. Fitting is a difficult problem where it is necessary to explore all possible angles and orientations of the atomic model that can fit the EM model. Typically, a single EM model is used for fitting. A statistical EM model, like ours, can provide a better range of fitting results where the atomic model can be fitted into one of several possible spike shapes, providing more flexibility. Such powerful modeling can further aid in providing more accurate atomic models of critical subcomponents of biological complexes such as the spike region of the AIDS virus.

2) Computer aided drug design: Development of computer algorithms for bio-simulations have enabled extensive studies of the protein structure and dynamics of new potential drug targets [21]. Numerous docking programs are extensively used in the biotechnology and pharmaceutical industries. The algorithms for docking use force-field-based methods such as molecular dynamics or Monte Carlo simulations which allow for movements of ligands and targets [21]. Most docking programs assume the structure to be rigid with a single shape. Using statistical shape models instead of a single shape model, docking and other computer aided drug design results could be further enhanced. Shape complementarity based methods fit the ligand shape into the negative shape of the protein structure. Using the statistical shape model of the virus spike, shape complementarity tests can be performed on candidate drugs using a wide range of possible spike shapes with associated statistical parameters.

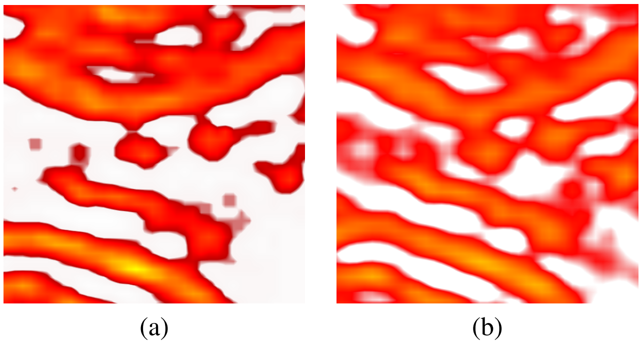


Fig. 14: Tomographic Reconstruction using (a) Shape based regularization, where the mean spike model was used as the prior shape in the reconstruction process, (b) Weighted Back Projection reconstruction. The shape based reconstruction method shows reduced blurring and improved spike feature visualization.

3) Shape based tomographic reconstruction: A recent tomographic reconstruction method introduced by Gopinath *et al.* [22] uses known shape models of certain critical structures in the tomographic reconstruction process. The shape models are incorporated into the tomographic reconstruction through a regularization process and a MAP (maximum *a posteriori*) estimate is obtained. Local segmentation of features is performed at each iteration of the reconstruction process and compared with the prior shape model. Over or under-segmentation detected at each voxel of the local feature drives a scaling factor of the regularization term.

As a simple example of the power of our derived spike shape model, we modified the reconstruction algorithm [22] using it as a shape prior. Specifically, the mean shape of the statistical spike model (Section III-B) was used as the prior shape in the regularization process. The resulting reconstruction shows reduced blur and improved feature visualization (Fig. 14). Further enhancements could an include a full Bayesian reconstruction scheme that completely utilizes the 4D statistical spike model in the reconstruction process as the prior probability distribution. Improved Electron Tomography reconstruction will provide better structure visualization giving important biological information around the vicinity of critical structures like the virus spikes.

IV. Conclusion

We introduced a fully automated technique to extract the spike features of the AIDS virus. Our method uses biological and structural information about the AIDS virus and the spike position and orientation vis-a-vis the virus to detect and extract these spikes. We used 3D volumetric images of the AIDS virus reconstructed from tilt series projection images generated from an electron microscope.

Our method is the first fully automated technique that can extract sub-volumes of the AIDS virus spike and build a statistical model without the need for any manual supervision. This is a significant improvement over current methods (Liu et al. [1], Zhu, et al. [3]) where biologists and biochemists use manual supervision to extract spikes and build a single average model. Our method can accelerate and increase the image data processing capacity of biochemists who seek to build models of the AIDS virus. Increased sample size as a result of larger data processing can lead to more accurate models of the virus spike. Shape complementarity between the spike and drug molecule is critical for the drug to effectively bind with the spike and neutralize the virus. Powerful statistical shape models can help in better drug design strategies. Currently our statistical model is based on 72 individual spike samples. With access to a larger set of electron microscopy data of the AIDS virus, we can build more powerful statistical models. Higher resolution data would also help, as the FP rate would be reduced leading to more ideal FROC curves. Using the tools developed for this method, we can analyze and build models of the AIDS virus envelope and other features of interest. Through minor modifications, our method can be easily extended to detect structures on the envelope of other virus and bacteria particles, which would be of interest for drug design.

References

[1] J. Liu, A. Bartesaghi, M. J. Borgnia, G. Sapiro, and S. Subramaniam, “Molecular architecture of native hiv-1 gp120 trimers,” *Nature*, vol. 455, pp. 109–113, September 2008.

[2] R. Narasimha, I. Aganj, A. E. Bennett, M. J. Borgnia, D. Zabransky, G. Sapiro, S. W. McLaughlin, J. L. Milne, and S. Subramaniam, “Evaluation of denoising algorithms for biological electron tomography,” *Journal of Structural Biology*, vol. 164, no. 1, pp. 7–17, 2008.

[3] P. Zhu, J. Liu, J. B. Jr, E. Chertova, J. D. Lifson, H. Grise, G. A. Ofek, K. A. Taylor, and K. H. Roux, “Distribution and three-dimensional structure of aids virus envelope spikes,” *Nature*, pp. 847–852, June 2006.

[4] T. Heimann and H.-P. Meinzer, “Statistical shape models for 3d medical image segmentation: A review,” *Medical Image Analysis*, vol. 13, pp. 543–563, 2009.

[5] F. L. Bookstein, *Morphometric tools for Landmark Data*. Cambridge University Press, 2003.

[6] S. M. Pizer, D. S. Fritsch, P. A. Yushkevich, V. E. Johnson, and E. L. Chaney, “Segmentation, registration and measurement of shape variation via image object shape,” *IEEE Transactions on Medical Imaging*, pp. 851–865, 1999.

[7] B. Tsagaan, A. Shimizu, H. Kobatake, and K. Miyakawa, “An automated segmentation method for kidney using statistical information,” in *Proc. MICCAI LNCS*, vol. 2488, 2002.

[8] T. Cootes and C. J. Taylor, “Combining point distribution models with shape models based on finite-element analysis,” *Image and Vision Computing*, vol. 13, no. 5, pp. 403–409, 1995.

[9] Z. Zhao, S. R. Aylward, and E. K. Teoh, “A novel 3d partitioned active shape model for segmentation of brain mr images,” in *Proc. MICCAI LNCS*, vol. 3749, Springer, 2005.

[10] D. Rueckert, A. F. Frangi, and J. A. Schnabel, “Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration,” *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 1014 – 1025, 2003.

[11] U. Grenander and M. I. Miller, “Computational anatomy: an emerging discipline,” *Quarterly of Applied Mathematics*, vol. LVI, pp. 617 – 694, 1998.

[12] U. Grenander and M. I. Miller, *Pattern Theory: From Representation to Inference*. Oxford University Press, 2007.

[13] M. I. Miller, “Computational anatomy: shape, growth, and atrophy comparison via diffeomorphisms,” *NeuroImage*, vol. 23, pp. S19–S33, September 2004.

[14] T. Rohlfing, R. Brandt, C. R. Maurer, and R. Menzel, “Bee brains, b-splines and computational democracy: Generating an average shape atlas,” in *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pp. 187–194, 2001.

[15] A. Bennett, J. Liu, D. V. Ryk, D. Bliss, J. Arthos, R. M. Henderson, and S. Subramaniam, “Cryoelectron tomographic analysis of an hiv-neutralizing protein and its complex with native viral gp120,” *The Journal of Biological Chemistry*, vol. 282, pp. 27754–27759, 2007.

[16] L. A. Shepp and Y. Vardi, “Maximum likelihood reconstruction for emission tomography,” *IEEE Transactions on Medical Imaging*, vol. MI-1, no. 2, pp. 113 – 122, 1982.

[17] R. Nürnberg, “Distance from a point to an ellipsoid,” tech. rep., Imperial College London, 2006.

[18] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, “Complex wavelet structural similarity: A new image similarity index,” *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2385 – 2401, 2009.

[19] D. P. Chakraborty and L. H. Winter, “Free-response methodology: alternate analysis and a new observer-performance experiment.,” *Radiology*, vol. 174, pp. 873–881, 1990.

[20] M. G. Rossmann, “Fitting atomic models into electron-microscopy maps,” *Biological Crystallography*, vol. D56, pp. 1341–1349, 2000.

[21] C. A. Taft, V. B. da Silva, and C. H. T. de Paula da Silva, “Current topics in computer-aided drug design,” *Journal of Pharmaceutical Sciences*, vol. 97, no. 3, pp. 1089–1098, 2008.

[22] A. Gopinath, G. Xu, D. Ress, O. Öktem, S. Subramaniam, and C. Bajaj, “Shape based regularization of electron tomography reconstruction,” *IEEE Transactions on Medical Imaging*, Submitted.