Generating a Statistical Shape Model of the AIDS Virus Spike

Ajay Gopinath and Alan C. Bovik

***Abstract*—We introduce a method to automatically extract the spike features of the AIDS virus imaged through an electron microscope. This method detects the location of the spikes and extracts a sub-volume enclosing the spike with a sensitivity of 80%. 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. This is a fully automated process that can extract sub-volumes of the AIDS virus spike making it possible to build a statistical model without the need for any user supervision. The AIDS virus spike is the primary target of drug design as it is directly involved in infecting host cells. 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.**

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

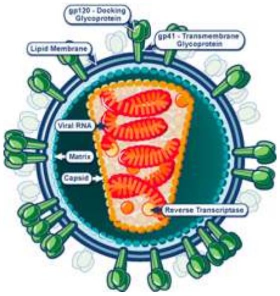

gp120 - Docking Glycoprotein
Lipid Membrane
gp41 - Transmembrane Glycoprotein
Viral RNA
Matrix
Capsid
Reverse Transcriptase

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)

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 $120\mathring{A}$ with a maximum width of about $150\mathring{A}$, tapering to $35\mathring{A}$ 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. 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.

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 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.

To build a statistical shape model of the AIDS virus spike, we use spike sub-volumes that are extracted and aligned automatically. We combine intensity information from all the individual detected spikes. The resulting statistical model of

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

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

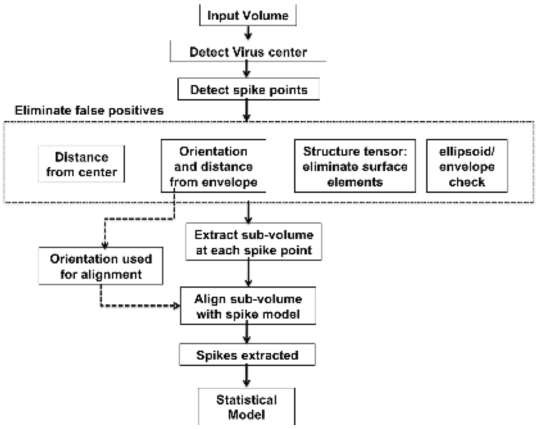

Input Volume
Detect Virus center
Detect spike points
Eliminate false positives
Distance from center
Orientation and distance from envelope
Structure tensor: eliminate surface elements
ellipsoid/ envelope check
Extract sub-volume at each spike point
Orientation used for alignment
Align sub-volume with 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 ±69° from a TEM. The images are of the Simian Immunodeficiency Virus (the HIV-like retrovirus that causes AIDS in monkeys) [8]. We used a Maximum Likelihood reconstruction scheme to perform the tomographic reconstruction [9]. Since projections from limited angles are available in Electron Tomography, it results in severe blurring of the biological structure being imaged, making the virus isolation, identification and modeling problems much more difficult. 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 hollow 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 preprocessing 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. The local maxima is calculated for each of the NCC volumes, which represents the virus center.

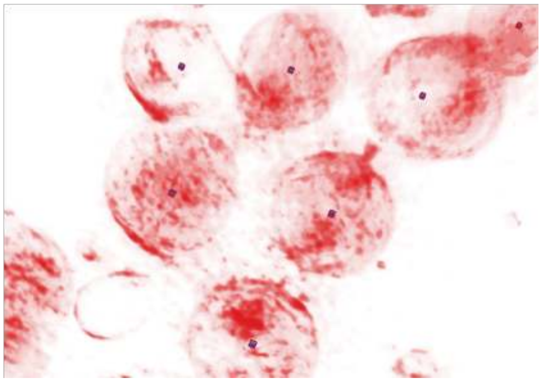


Fig. 3: 3D volume rendering of the virus particles with the detected virus centers depicted as dots inside.

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*. 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). Local maxima are located for each DoG volume and its immediate neighbors. The next steps attempt to eliminate the false positives and preserve only those points that lie on a spike.

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 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

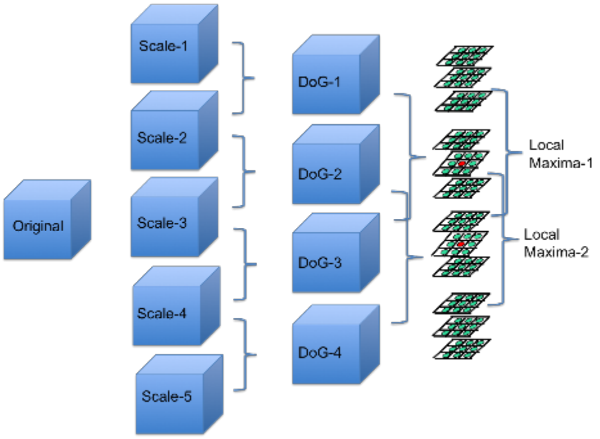


Fig. 4: Detecting spike-points: The subvolume containing the virus particle is scaled by convolution with a Gaussian kernel at different sigmas. Difference of Gaussian (DoG) volumes are computed by subtracting neighboring scaled volumes. Local maxima of each DoG volume including the local neighborhood 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.

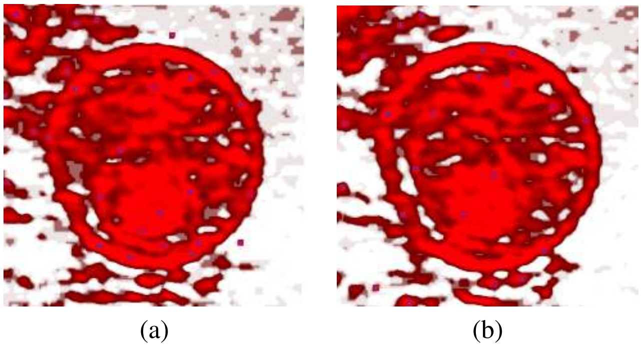

(a) (b)

Fig. 5: Spike-points before FP removal: (a) and (b) are 2D slices of a virus particle with spike-points shown in blue.

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. 6). The virus center detection step (Section II-A1) 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, which is the predicted spike orientation. The distance from the spike point to the envelope along the normal provides another check for false positive elimination.

c) Structure tensor: 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.

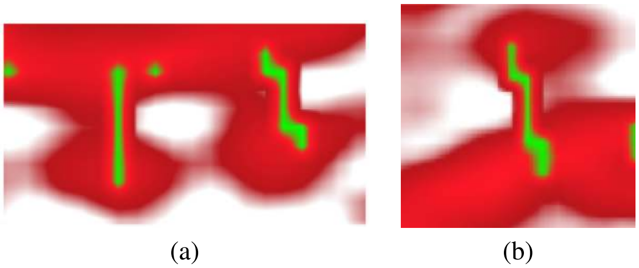

(a) (b)

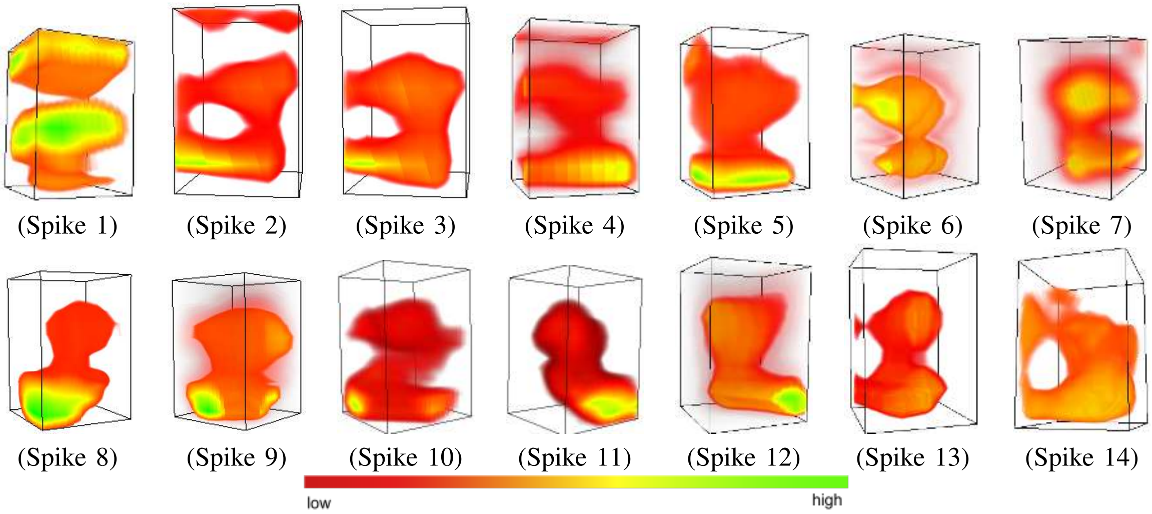
Fig. 6: Spike orientation axis: (a, b) are 2D slices with the detected spike orientation axis shown in green.

4) Extracting spikes: 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 a Free response Receiver Operating Characteristics analysis. At high threshold values, spike points 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 compared with a phantom spike model to recover its orientation.

This procedure was performed for all the spike points and we extracted the sub-volume containing the spike and also determined its orientation (Fig. 7). With the orientation recovered, it is now possible to combine all the extracted spikes in the next step of building a statistical shape model.

III. Results and discussion

We measured 80% sensitivity (80% of all spikes were detected) where 77 out of 96 spikes were detected with 9 false positives (FP) per virus. Our best operating range is at 7 FP with a sensitivity around 75%, beyond which the sensitivity drops. The statistical model is in 4D data where the fourth dimension represents the probability density function associated with each voxel. Figs. 8 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. 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 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. 8, 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.


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

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

Fig. 7: 3D Volume rendering of automatically extracted spikes aligned to a common coordinate frame.

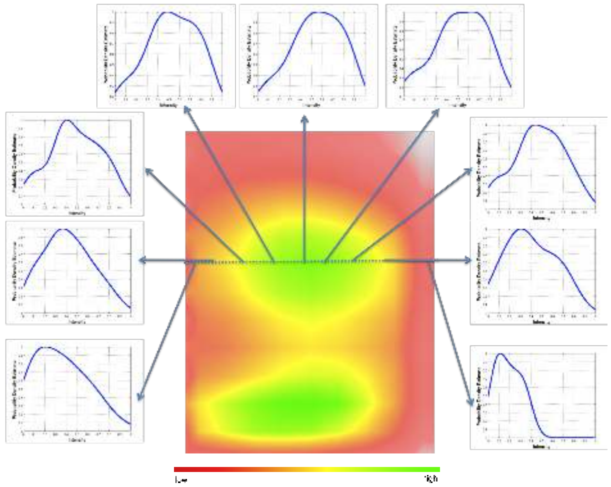


Fig. 8: Statistical Shape Model: A 2D profile image of the mean statistical model along the XZ plane and plots of the density function associated with voxels that lie on the center line shown in the image.

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 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. 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] 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.

[9] 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.