NANO-SCALE CONTEXT-SENSITIVE SEMANTIC SEGMENTATION

Nan Zhao and Xiuwen Liu

Florida State University
Department of Computer Science
Tallahassee, FL, 32310

ABSTRACT

Nano-scale imaging technologies make it possible to visualize objects at nanometer resolutions. To investigate structures and functions of interest, there is an intrinsic demand for explicit models to extract them from nano-scale data. Segmentation is one of the most critical steps in processing pipelines. However, existing segmentation methods often fail due to extremely low signal-to-noise ratio, low contrast and large data size. In this paper we propose a new context-sensitive method for segmenting three-dimensional volumes. As our method efficiently narrows the search space by using robust context cues, we achieve tractable and reliable nano-scale semantic segmentation. We demonstrate our method on a tomogram of microvilli spikes, for which our method is able to yield accurate spike segmentation and in comparison the state-of-the-art semantic segmentation methods fail due to their inability to handle signal-to-noise ratio and low contrast volumes.

***Index Terms*—** Nano-scale, context-sensitive, semantic segmentation, microvilli, spike

1. INTRODUCTION

Nano-scale imaging technologies allow us to investigate nano-scale structures that are very close to their native states and potential spatial relations between them. Thus they benefit a wide range of nano-scale studies such as biochemistry, material science, and medicine [1, 2]. Generally, nano-scale structure studies rely on several critical stages, among which target object identification and segmentation are of utmost importance [3]. In the literature of computer vision, a task combining both problems is called *semantic segmentation*. The common and natural way of nano-scale semantic segmentation is carried out manually, using some visualization tools [4, 5]. This is not only subjective but also labor-expensive, especially when the rapid advances in automation of nano-scale imaging have led to a dramatic increase in the speed of data collection. Therefore, the growing amount of human effort required in segmentation becomes the bottleneck of nano-scale researches.

There exists a large number of segmentation algorithms that attempt to automate the segmentation process, represented by watershed [6], active contour [7, 8], level set [9], sliding window [10], GraphCut [11] and Gaussian Mixture Model (GMM) [12]. However, these algorithms achieve limited success in nano-scale data [1, 3]. The task of automation on nano-scale segmentation is primarily hindered by the coexistence of two problems: the low signal-to-noise (SNR) ratio and the large scale (i.e.: a $600 \times 1400 \times 432$ tomogram). The first problem is intrinsic to nano-scale imaging because of how the image is produced. In an attempt to imaging nano-scale objects, it is necessary to use enough doses of electrons to capture measurable contrast. On the other hand, an increase in the use of electron does tends to damage the structure of nano-scale objects. Based on this trade-off, it is common to observe nano-scale data with low SNR and low contrast (as shown in Fig. 1 (a)). The second problem derives from the first problem. Due to the low SNR, it is often very difficult to identify the target objects (such as the spikes in Fig. 1 (a)) without the existence of much larger context objects (such as the membranes in Fig. 1 (b)). Thus high resolution is often necessary for capturing both small "*context sensitive*" target objects and large context objects. Consequently, all the methods above become intractable and inapplicable to nano-scale semantic segmentation – a task of small object segmentation in big and noisy data.

To address these issues, we assume the context object segmentation are tractable using appearance features and then propose a two-stage framework, named *context-sensitive semantic segmentation*. The key idea is that, via the cues of context objects that are robust to the noise, our framework efficiently narrows the search space of the target object identification in the high resolution data. As a test case, we apply our segmentation framework on a tomogram of microvilli with membranes and spikes, although clearly it is not limited to this example. Our main contributions are as follows:

- We propose a novel statistic framework that can be extensively used with different context features to overcome the problems of nano-scale semantic segmentation (Section 2).
- We design effective context features to achieve semi-automatic microvilli spike segmentation (Section 3).

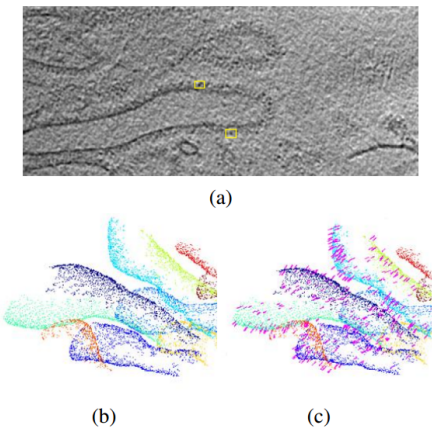

Fig. 1. Our task is to address small and faint object (in our example, spikes) segmentation on a nano-scale tomogram. An exemplar "slice" of a tomogram is shown in (a), with two sample spikes marked by yellow squares. In the first stage, we produce context object segmentation (membranes colored in (b)). Each color indicates one connected component in 3D. In the second stage, we model a number of context cues and propose a hybrid model to produce a voxel-wise segmentation of target objects (3D spike ridges in (c) with magenta labels).

2. FRAMEWORK

2.1. Classical semantic segmentation framework

In a general statistic framework, 3D semantic segmentation is modeled as finding the label o_i of each voxel i that maximizes $\Pr(o_i|f_i)$, the conditional probability density function (PDF) of the presence of the object o_i given a set of features f_i. In classical framework for semantic segmentation, objects in the background are considered as noise, rather than cues. Hence, it is often called *object-centered model*. Respectively, we have $\Pr(o_i|f_i) \simeq \Pr(o_i|f_i^A)$, where f_i^A is a set of local appearance features of the target object. Unfortunately, the assumption of the object-centered framework does not often hold in nano-scale. The intrinsic object appearance features are often not distinctive enough for accurate semantic segmentation when SNR is extremely low. Another drawback is its computational cost. Note that f_i^A is a feature set, every feature needs to be generated through measurements across different locations and scales of the entire volume. Thus the scalability of this framework is intrinsically limited by the large searching space.

2.2. Our two-stage context-sensitive semantic segmentation framework

Instead of modeling context objects in the background as noise, we propose context-sensitive semantic segmentation – a new framework that is sensitive to contextual features provided by objects in the background, namely *context objects*. The problem of semantic segmentation on faint object in nano-scale is then re-modeled as two stages: semi-global context object segmentation and faint target object segmentation.

2.2.1. Stage one: context object segmentation

As we assume appearance features of context objects are distinctive enough to produce segmentation, $\Pr(o_i|f_i)$ can be written as $\Pr(o_i|f_i^A)$ again, where f_i^A is a set of appearance features of context object in voxel i, $o_i = 1$ means semi-global context object and 0 otherwise. For simplicity, we model this stage as a binary segmentation problem using just thresholding, which could be replaced by more sophisticated segmentation algorithms depending on varied demands [3]. Note that f_i^A could be any specific object features in any specific problem, relying on which objects provide the contextual information.

2.2.2. Stage two: faint target segmentation

The availability of hard segmentation on context objects $O = \{o_i\}_{i=1}^N$ allows us to compute the appearance and contextual features ($f_i^{A'}$ and $f_i^{C'}$ respectively) for faint and small target objects. The second stage is thus modeled as finding a discriminant function $\Pr(o_i{}' = 1|f_i^{A'}, f_i^{C'})$ that predicts the posterior probability of a faint target at the $i'th$ voxel given both types of features. Here $f_i^{A'}$ and $f_i^{C'}$ are used to summarize all types of appearance and contextual features for faint target at the $i'th$ voxel, $o_i{}' = 1$ means target object and 0 otherwise. After factorization we have

$$\begin{aligned} &\Pr(o_i' = 1|f_i^{A'}, f_i^{C'}) \\ &\propto \Pr(o_i' = 1|f_i^{A'}) \times \Pr(f_i^{C'}|o_i' = 1). \end{aligned} \tag{1}$$

The first term of (1) is simply the classical object-centered model. The second term is a log-likelihood term that favors context feature responses that are consistent with our prior knowledge about the target. For instance, if it is known that the targets are cars, then the log-likelihood term will be much larger for road regions than for sea regions.

Following the survey of contexts [13], we extend $f_i^{C'}$ into three sets– the semantic context $f_i^{C_{se}}$(e.g: probability of co-existence), the spatial context $f_i^{C_{sp}}$(e.g.: position and orientation) and the scale context $f_i^{C_{sc}}$(e.g.: size) of the target object with respect to its nearby context object respectively. Eq. (1) can thus be decomposed into four terms,

$$\begin{aligned} &\Pr(o_i' = 1|f_i^{A'}, f_i^{C'}) \propto \Pr(o_i' = 1|f_i^{A'}) \times \Pr(f_i^{C_{se}}|o_i' = 1) \\ &\times \Pr(f_i^{C_{sc}}|f_i^{C_{se}}, o_i' = 1) \times \Pr(f_i^{C_{sp}}|f_i^{C_{sc}}, f_i^{C_{se}}, o_i' = 1), \end{aligned} \tag{2}$$

where each of the last three terms takes an additional type of contextual cues into account sequentially. In our work, we focus on how to utilize different types of contextual cues to not

only improve the segmentation accuracy but also significantly accelerate the segmentation on large scale.

3. IMPLEMENTATION

3.1. Context object segmentation

The assumption of our framework is that the appearance features for larger-scale context object are distinctive enough for segmentation. Here we just follow the method of Antonio etc. [14] and extract the context objects – the microvilli membranes $M = \{M_k\}$. Fig. 1(b) illustrates a 3D view of the extracted microvilli membranes.

3.2. Fine-scale faint target segmentation

3.2.1. Appearance cue

As spike heads are somewhat darker than the local background in certain scale space, they appear as local minimums in the 3D tomogram. Thus we smooth the image I with a 2D Gaussian filter G of variance σ', $H = I * G_{\sigma'}$, and then generate the appearance model:

$$\Pr(o_i{}' = 1|f_i^{A'}) = \begin{cases} \psi(H_i) & \text{, if } i = \arg\max_{j\in\mathcal{N}_i} H_j, \\ 0 & \text{, otherwise,} \end{cases} \tag{3}$$

such that $\psi(H_i) = (\max(H) - H_i)/(\max(H) - \min(H))$ and $\mathcal{N}_i$ is the voxel i with its 26 neighbor voxels in 3D.

3.2.2. Scale context cue

Let t and h be the thickness of the membrane and the maximum length of a spike respectively. Given the membrane mask M, we form a number of zones to exclude the local minimums due to not only membranes but also background noises that are far from the membrane:

$$f_i^{C_{sc}} = M \oplus E_h - M \oplus E_t, \tag{4}$$

where $\oplus$ denotes the 3D morphological dilation. Correspondingly, the likelihood of scale context feature is as follows:

$$\Pr(f_i^{C_{sc}}|f_i^{C_{se}}, o_i{}' = 1) = \begin{cases} f_i^{C_{se}}, & \text{if } f_i^{C_{sc}} = 1, \\ 1 - f_i^{C_{se}}, & \text{otherwise,} \end{cases} \tag{5}$$

where $f_i^{C_{se}} = 1$ if the root of the spike that contains voxel i is labeled as 1 in membrane segmentation result M, whereas $f_i^{C_{se}} = 0$ if the respective root is labeled as 0 in M. Hence, we formulate the fact that the scale context cue must be satisfied ($f_i^{C_{sc}} = 1$) if the labels of the target (the spike head) and the context (the respective spike root on the membrane) are both given.

3.2.3. Spatial context cue

Due to the prior knowledge that spikes are perpendicular to and grow toward the outside of the membrane while the shape of the membrane is convex in general, the root of each spike should be closer to the membrane centroid than the spike head. Let $d(.,.)$ be the euclidean distance between two voxels given their indexes. We can compute the spatial context feature as:

$$f_i^{C_{sp}} = \frac{d(c_k, i)}{d(c_k, i'_M)}, \tag{6}$$

where c_k is the centroid index of membrane mask M_k and i'_M is the index of the spike root corresponding to the potential spike head i on membrane segmentation M. As $f_i^{C_{sp}}$ is larger than 1 if the voxel i is outside the membrane, the likelihood of spatial context feature is as follows:

$$\Pr(f_i^{C_{sp}}|f_i^{C_{se}}, o_i{}' = 1) = \begin{cases} f_i^{C_{se}}, & \text{if } f_i^{C_{sp}} > 1, \\ 1 - f_i^{C_{se}}, & \text{otherwise.} \end{cases} \tag{7}$$

Similarly, here we model the spatial context cue that a spike root must be closer to the center of the arrayed membrane than its respective spike head is, indicating the outside of the membrane.

3.2.4. Semantic context cue

Due to the missing wedge effect of nano-scale imaging, both membranes and spikes are partially blurred or even missed. Thus it is necessary to take the reliability of membrane segmentation into account. We explicitly model the semantic context cue as the coefficient in a hybrid model, determining the relative contribution of appearance features and context features in semantic segmentation:

$$\Pr(f_i^{C_{se}}|o'_i = 1) = \begin{cases} \lambda, & \text{if } f_i^{C_{se}} = 1, \\ 1 - \lambda, & \text{if } f_i^{C_{se}} = 0. \end{cases} \tag{8}$$

If we assume the scale context feature $f_i^{C_{sc}}$ and the spatial context feature $f_i^{C_{sp}}$ are conditionally independent of each other given the semantic context feature and the target label, the spike likelihood channel, Eq.(2), is further modeled as:

$$\Pr(o'_i = 1|f_i^{A'}, f_i^{C_{sc}}, f_i^{C_{sp}}, f_i^{C_{se}}) \propto \lambda\Psi^C + (1-\lambda)\Psi^A,$$

such that

$$\begin{aligned} \Psi^C &= \Pr(o'_i = 1|f_i^{A'}) \times \Pr(f_i^{C_{sc}}|f_i^{C_{se}} = 1, o'_i = 1) \\ &\quad \times \Pr(f_i^{C_{sp}}|f_i^{C_{se}} = 1, o'_i = 1), \\ \Psi^A &= \Pr(o'_i = 1|f_i^{A'}). \end{aligned} \tag{9}$$

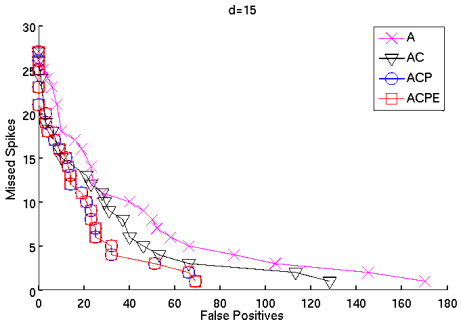

Fig. 2. Spike head segmentation performance of different models. See the text for a description of each model.

The semantic context is explicitly formulated by λ. Then the classic object-centered model is a special case of our model when $\lambda = 0$. When λ increases, more contributions from the context cues are taken into account in semantic segmentation and the segmentation is thus more sensitive to the context. When $\lambda = 1$, our model is close to the context integration models ([15, 16, 17]).

4. EXPERIMENTAL RESULTS

We evaluated the performance of our method on spike segmentation in a $600 \times 1400 \times 432$ tomogram acquired from the microvilli of insect flight muscle. 27 spike heads arrayed on a membrane within two slices were annotated by a microvillus expert. Thus we formulate the evaluation of spike segmentation as evaluating the spike head detection and follow the evaluation methodology of the PASCAL object detection challenges [18]. A detected voxel (thresholding on Eq. 9) and a groundtruth voxel form a match if their euclidean distance is smaller than $d = 15$, which is suggested as the minimum distance between two spikes. Moreover, the low SNR and the large size yield difficulty in using most up-to-date segmentation methods as a good baseline technique for comparison. As our task is to show the contribution of context cues in our framework, we apply thresholding on Eq. 3 as a baseline algorithm that is purely based on the appearance feature.

In Figure 2, we present the performance of different models in terms of the number of missed spikes against the number of false positives. Model names are shown as follows: 'A' is our baseline object-centered model, 'ACPE' is our complete model, and 'AC' and 'ACP' are ablations of our complete model. Here 'A', 'C', 'P', 'E' implies appearance feature, scale context feature, spatial context feature and semantic context feature respectively. In Fig. 3, we visualized an exemplar slice cut of several sample output spikes of several models ($\lambda = 0.8$), along with the ground truth annotation in the cropped original tomogram. By projecting the detected spike heads of ACPE when FP=50 onto their closest membranes, we have the corresponding spike ridges shown in Fig. 1(c). Our results have shown the helpfulness of different context cues in our task clearly. The baseline algorithm does a worse job than the other algorithms with context cue(s). Moreover, our complete model reaches the best performance. Even though the contribution of semantic context hasn't been shown, it completes our general model so that our model allows the use of all types of potentially useful context cues.

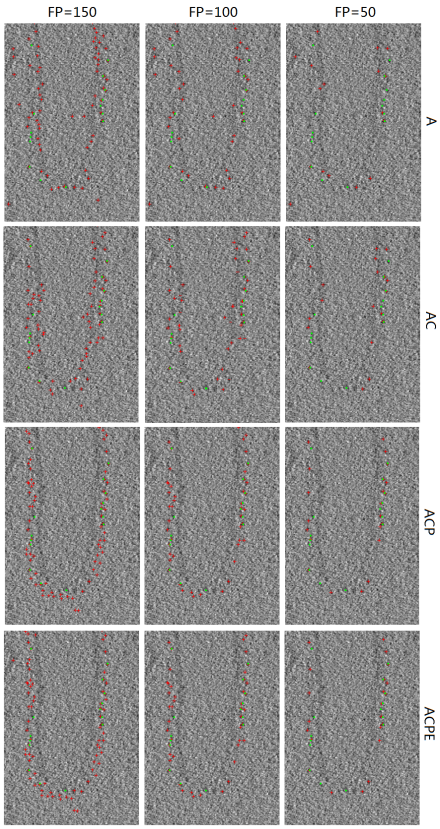

Fig. 3. Visualization of spike head segmentation on a exemplar slice of the tomogram. The green dots are the ground truth. The red crosses are the spike heads detected by the respective model. See the text for a description of each model.

5. CONCLUSION

We have presented a novel framework for nano-scale semantic segmentation, demonstrated on spike segmentation in a cryo-electron tomogram. The low SNR and high resolution of our data makes most up-to-date segmentation methods intractable. In contrast, our method achieved efficient voxel-wise segmentation through context features that do not only tolerate the extremely noisy background, but also reduce the searching space dramatically.

6. REFERENCES

[1] Eraldo Ribeiro and Mubarak Shah, "Computer vision for nanoscale imaging," *Machine Vision and Applica-*

tions, vol. 17, no. 3, pp. 147–162, 2006.

[2] Ping Zhu, Jun Liu, Julian Bess, Elena Chertova, Jeffrey D Lifson, Henry Grisé, Gilad A Ofek, Kenneth A Taylor, and Kenneth H Roux, "Distribution and three-dimensional structure of aids virus envelope spikes," *Nature*, vol. 441, no. 7095, pp. 847–852, 2006.

[3] Jose-Jesus Fernandez, "Computational methods for electron tomography," *Micron*, vol. 43, no. 10, pp. 1010–1030, 2012.

[4] James R Kremer, David N Mastronarde, and J Richard McIntosh, "Computer visualization of three-dimensional image data using imod," *Journal of structural biology*, vol. 116, no. 1, pp. 71–76, 1996.

[5] Benjamin Schmid, Johannes Schindelin, Albert Cardona, Mark Longair, and Martin Heisenberg, "A high-level 3d visualization api for java and imagej," *BMC bioinformatics*, vol. 11, no. 1, pp. 274, 2010.

[6] Luc Vincent and Pierre Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 13, no. 6, pp. 583–598, 1991.

[7] Andrew Blake and Michael Isard, *Active contours: the application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion*, Springer-Verlag New York, Inc., 1998.

[8] Alberto Bartesaghi, Guillermo Sapiro, and Sriram Subramaniam, "An energy-based three-dimensional segmentation approach for the quantitative interpretation of electron tomograms," *Image Processing, IEEE Transactions on*, vol. 14, no. 9, pp. 1314–1323, 2005.

[9] Daniel Cremers, Mikael Rousson, and Rachid Deriche, "A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape," *International journal of computer vision*, vol. 72, no. 2, pp. 195–215, 2007.

[10] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[11] Yuri Boykov and Gareth Funka-Lea, "Graph cuts and efficient nd image segmentation," *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006.

[12] Yi Ma, Harm Derksen, Wei Hong, and John Wright, "Segmentation of multivariate mixed data via lossy data coding and compression," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 9, pp. 1546–1562, 2007.

[13] Carolina Galleguillos and Serge Belongie, "Context based object categorization: A critical survey," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 712–722, 2010.

[14] Antonio Martinez-Sanchez, Inmaculada Garcia, and Jose-Jesus Fernandez, "A differential structure approach to membrane segmentation in electron tomography," *Journal of structural biology*, vol. 175, no. 3, pp. 372–383, 2011.

[15] Antonio Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[16] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Computer Vision–ECCV 2006*, pp. 1–15. Springer, 2006.

[17] Zheng Song, Qiang Chen, Zhongyang Huang, Yang Hua, and Shuicheng Yan, "Contextualizing object detection and classification," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1585–1592.

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.