doi:10.1017/S143192761801320X
Microsc. Microanal. 24 (Suppl 2), 2018

A Pipeline for Distributed Segmentation of Teravoxel Tomography Datasets

Mehdi Tondravi[1], William Scullin[2], Ming Du[3], Rafael Vescovi[4], Vincent De Andrade[4], Chris Jacobsen[4,5,1], Konrad Paul Kording[6], Doğa Gürsoy[4,7], and Eva Dyer[8,*]

[1.] Chemistry of Life Processes Institute, Northwestern University, Evanston, IL USA
[2.] Argonne Leadership Computing Facility, Argonne National Lab, Lemont, IL USA
[3.] Dept. Materials Science, Northwestern University, Evanston, IL USA
[4.] Advanced Photon Source, Argonne National Lab, Lemont, IL USA
[5.] Dept. Physics & Astronomy, Northwestern University, Evanston, IL USA
[6.] Dept. Bioengineering and Neuroscience, University of Pennsylvania, Philadelphia, USA
[7.] Dept. Electrical Engineering and Computer Sci., Northwestern University, Evanston, IL USA
[8.] Dept. Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA USA
* Corresponding author, evadyer@gatech.edu

X rays are able to penetrate specimens that are centimeters thick, while also delivering tomography data with sub-micrometer voxel size. This leads to teravoxel datasets from which one would like to obtain a segmented representation to identify and analyze various features. While there are a number of toolkits available for segmentation on gigavoxel datasets that fit within the memory of a single computer workstation, for larger datasets one must turn to parallelized segmentation on distributed computing systems.

Here, we introduce a distributed data processing pipeline for segmenting teravoxel (and larger) image volumes. Our workflow (shown in Fig. 1) involves the splitting of a 3D volume data array into a large number of small sub-volumes with overlap, applying a range of classifiers to each of the subarrays to identify the features of interest, and then merging the outputs of the classifier to obtain a final result on the full 3D volume. To classify each sub-volume, we use *ilastik* [1], an existing open-source toolset that allows the user to interactively annotate a small set of image slices as training for segmentation of larger volumes. In our previous work, we have shown that *ilastik* provides a user friendly front-end for robust classification that can be used to segment millimeter-sized samples of mouse brain cortex [2]. Our results suggest that this workflow based on *ilastik* can be extended to significantly accelerate processing of large-scale X-ray tomography datasets.

While the proposed data processing pipeline can be used in a wide range of distributed computing environments, we implemented it on the Cooley cluster of the Argonne Leadership Computing Facility (ALCF), which is available via no-cost scientific user proposals. The Cooley cluster has 126 compute nodes, each with two 2.4 GH Intel E5-2620 v3 processors with 6 cores per processor, two NVidia K80 GPUs, and 384 GB of RAM per node. Depending on data size, different number of parallelized processes (ranks) can be allocated for various steps of the workflow. The timings for these operations versus data size are shown in Fig. 2 (left), which shows near-ideal linear scalability. The code is publicly available (github.com/xbrainmap).

To demonstrate the utility of our pipeline, we applied it to a publicly available X-ray microtomography dataset (github.com/nerdslab/xbrain) collected from a cubic-mm brain sample from mouse somatosensory cortex [2]. The size of this entire dataset is 2560x2560x1624 (10.6 Gigavoxels). To generate a segmented output with our methods, we distributed our workflow on

48 Cooley-cores to produce an output in one hour including I/O. Our results demonstrate that this workflow can be used to obtain accurate classification and segmentation results (Fig. 2, right) with nearly linear scaling as we increase the number of cores (Fig. 2, left).

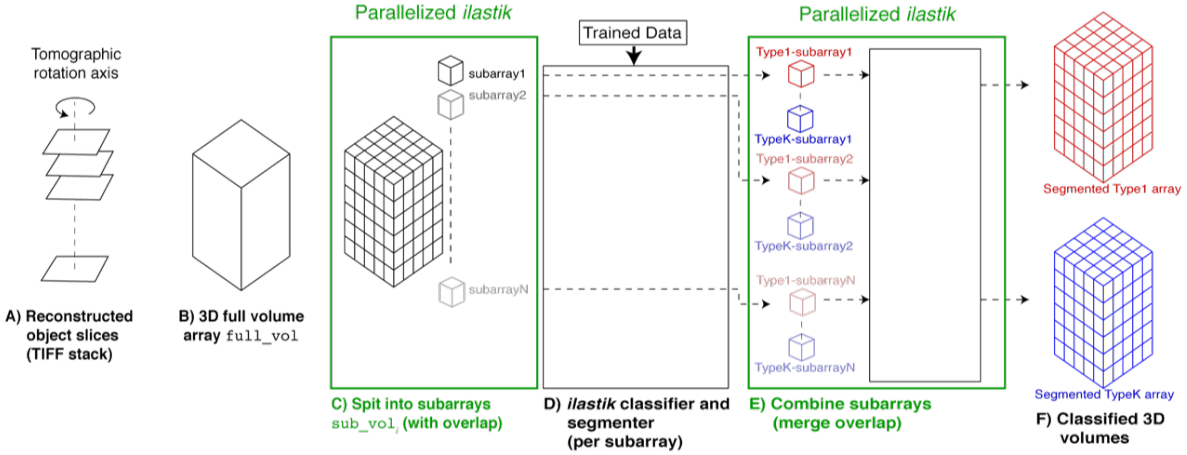

Figure 1. A 3D volume is divided into a number of overlapping subarrays, each subarray is processed by a single core, and then merged to yield full volume arrays of each of the classified features.

Quantity	4× downsampled			2× downsampled			Full resolution
N_z	4,600			9,200			18,000
N_y	2,900			5,800			11,600
N_x	4,100			8,200			16,400
Teravoxels	0.055			0.438			3.42
Rank r_{s-c} for splitting and combining	48			36			96
Nodes for splitting and combining	4			6			8
Splitting core-hours	6.4			18			110
Combining core-hours	35.2			164			365
Rank r_{seg} for segmentation	64	32	16	64	32	16	64
Cores for segmentation	768	384	192	768	384	192	768
Segmentation core-hours	368	330	310	2,670	2,520	2,460	21,696
Total core-hours	410	371	352	2,850	2,700	2,650	22,171
Core-hours per teravoxel	8,230	7,460	7,070	7,160	6,790	6,650	6,483
Total clock time (hours)	1.3	1.7	2.5	6.0	9.1	15	33

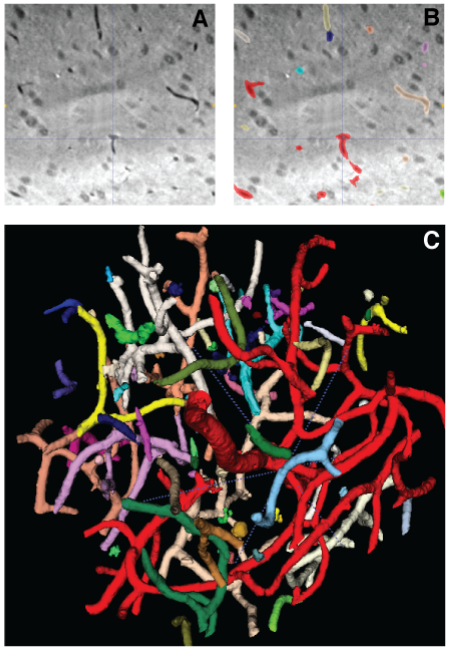

Figure 2. (Left) Timings for segmentation of a larger dataset (classifying x-ray tomography data of a mouse brain segmented for cell nuclei, blood vessels, and "other") at three different resolutions. (Right) Segmentation results on a small (400 x 400 x 400) brain volume imaged at 0.65 isotropic micron resolution: (A) original X-ray image, (B) segmented vessels overlaid), (C) 3D visualization of (B).

References:

[1] C. Sommer *et al*, in "2011 8th IEEE International Symposium on Biomedical Imaging (ISBI 2011)," (IEEE, New York), p. 230. See also `www.ilastik.org`.

[2] E. Dyer *et al.*, eNeuro **4** (2017), E0195-17.2017.

This work was funded by the National Institutes of Health under grant U01 MH109100, and U.S. Department of Energy, Office of Science, Offices of Advanced Scientific Computing Research and Basic Energy Sciences via Contract No. DE-AC02-06CH11357.

https://doi.org/10.1017/S143192761801320X Published online by Cambridge University Press