Ground Truth Estimation by Maximizing Topological Agreements in Electron Microscopy Data

Huei-Fang Yang* and Yoonsuck Choe

Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843-3112

Abstract. Manual editing can correct segmentation errors produced by automated segmentation algorithms, but it also introduces a practical challenge: the combination of multiple users' annotations of an image to obtain an estimation of the true, unknown labeling. Current estimation methods are not suited for electron microscopy (EM) images because they typically do not take into account topological correctness of a segmentation that can be critical in EM analysis. This paper presents a ground truth estimation method for EM images. Taking a collection of alternative segmentations, the algorithm seeks an estimated segmentation that is topologically equivalent and geometrically similar to the true, unknown segmentation. To this end, utilizing warping error as the evaluation metric, which measures topological disagreements between segmentations, the algorithm iteratively modifies the topology of an estimated segmentation to minimize the topological disagreements between this estimated segmentation and the given segmentations. Our experimental results obtained using EM images with densely packed cells demonstrate that the proposed method is superior to majority voting and STAPLE commonly used for combining multiple segmentation results.

1 Introduction

Electron microscopy (EM) image segmentation is the first step toward the reconstruction of neural circuits. However, because EM images show high variations in neuronal shapes and have ambiguities in boundary localization due to imperfect staining and imaging noise, automated segmentation algorithms sometimes generate incorrect results. Such errors require manual correction, that is, manual correction of erroneous segmentations is an important part of the neural circuit reconstruction pipeline.

The user's editing or correction improves segmentation accuracy, but it introduces a new practical challenge: the combination of multiple users' annotations of an image to obtain an estimation of the true, unknown labeling. Several combination methods have been proposed in the literature. The simplest combination

* This work was supported in part by NIH/NINDS #1R01-NS54252 and NSF CRCNS #0905041.

G. Bebis et al. (Eds.): ISVC 2011, Part I, LNCS 6938, pp. 371–380, 2011.

strategy is majority voting. It treats each individual segmentation equally and assigns a pixel the label that most segmentations agree on. Another commonly used combination method is to weight each segmentation differently according to the performance of each user. This method is referred to as global weighted voting. Simultaneous truth and performance level estimation (STAPLE) algorithm proposed by Warfield et al. [1] belongs to this category. STAPLE uses an iterative expectation-maximization algorithm to measure the performance of experts and estimates the underlying true segmentation by optimally combining each segmentation depending on each expert's performance level. In contrast to giving the same weighting to each pixel in a segmentation, local weighted voting methods [2,3] assign each pixel a different weight according to a local estimation of segmentation performance. However, these combination strategies are not adequate for the EM images because they do not take into account the topological or morphological correctness of a segmentation. Ensuring that a segmentation is topologically correct is a necessity to obtain an accurate neural circuit reconstruction. As pointed out by Jain et al. [4], in EM segmentation for reconstructing detailed connections between neurons, a topological error (i.e. merge or split errors) occurring at a branch causes severely erroneous neural connectivities in the subsequent sub-tree. Therefore, combination strategies for EM images should ensure that the estimate of the actual labeling given a set of segmentations is topologically correct.

This paper presents a segmentation ground truth estimation method for EM images from brain tissue. Taking a collection of annotations of an image, the algorithm aims at providing an estimated labeling that is topologically equivalent and geometrically similar to the true, unknown segmentation. To this end, guided by the segmentation evaluation metric, warping error, the algorithm iteratively modifies topology of the estimated segmentation to maximize the topological agreements (i.e. minimize the disagreements) between the estimated segmentation and a set of given segmentations. Topological change can be done by merging two adjacent regions or splitting a region into two by modifying the label of a sequence of pixels. By gradually changing its topology, the estimated segmentation becomes topologically equivalent to the true, unknown segmentation.

2 Evaluation Metric: Warping Error

In supervised evaluation, the performance of a segmentation algorithm is quantitatively measured by comparing its segmentation results against a manually labeled ground truth (i.e. a reference image) based on evaluation metrics. Jaccard index [6], Dice Similarity Coefficient (DSC), and F-measure are well known and widely used metrics for segmentation evaluation. Those metrics use the amount of overlap between a segmentation and the ground truth as a similarity measure to evaluate the performance of a segmentation method. This makes them focus on measuring a segmentation's boundary accuracy at the pixel level but not take into account its topological correctness. However, in EM segmentation evaluation, measuring the degree of the topological correctness of a segmentation is

also important because obtaining accurate analysis of the neural circuits relies on topologically correct reconstructions [5].

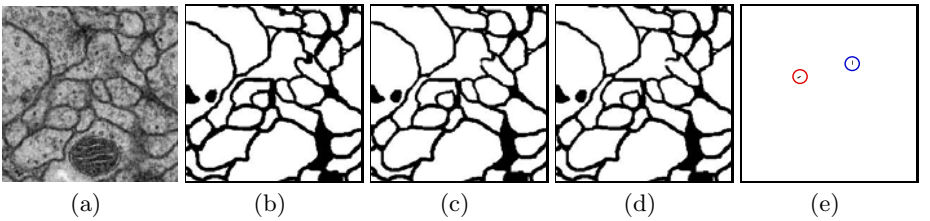
The warping error metric proposed by Jain et al. [5] is a metric that measures topological disagreements between segmentations and has been shown to be effective for EM segmentation evaluation. While comparing two segmentations, the error metric strongly penalizes topological disagreements but tolerates minor boundary localization differences. Conceptually, to calculate the topological disagreements between a segmentation and the ground truth, the ground truth image is first transformed into another image under topological and geometrical constraints, and then the disagreements (i.e. topological errors) can be identified as the pixel differences between the transformed image and the segmentation to be evaluated.

Before giving a formal definition of warping error, the concept of *warping* is first presented. Formally, given two binary images, L^* and L, if L^* can be transformed into L by flipping the labels of a sequence of pixels, L is called a *warping* of L^*, represented as $L \triangleleft L^*$. That is to say, L^* and L are topologically equivalent and geometrically similar. The labels of a sequence of pixels to be flipped are those of simple points (i.e. border pixels), where a point p is defined as

Algorithm 1. A warping algorithm that warps a binary image L^* to a segmentation T. Algorithm from [5].

input : A binary image L^*, a segmentation T, and geometric constraints set G
output: A warped image L

```
L = L*;
while true do
    S = simple(L) ∩ G ;
    i = argmax_{j∈S} |t_j − l_j|;
    if |t_i − l_i| > 0.5 then
        l_i = 1 − l_i;
    else
        return L;
```



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

Fig. 1. Measured warping error of a segmentation against the ground truth. (a) A sample EM image. (b) Manually annotated ground truth. (c) A warping of the ground truth image shown in (b) given a segmentation in (d). (d) A segmentation to be evaluated. (e) Measured warping error. The pixel disagreements between (c) and (d) consist of the topological errors. In this case, two topological errors, a merge (blue circle) and a split (red circle), occur due to the boundary ambiguity in the original image.

a simple point if both the number of foreground connected components adjacent to p and the number of background connected components adjacent to p equal to 1. According to the theory of digital topology, flipping the labels of simple points will not alter the object's topology.

Now, letting T be a segmentation to be evaluated and L^* be the reference annotation, the warping error $D(T \parallel L^*)$ is given as

$$D(T \parallel L^*) = \min_{L \triangleleft L^*} |E(L,T)|, \tag{1}$$

where L is the optimal warping of L^*, and $E(L,T)$ is the difference set (i.e. the pixels that have different labels in images L and T), defined as $E(L,T) = L \triangle T$. In other words, the warping error is considered as the pixel disagreements between the segmentation to be evaluated T and the transformed segmentation L. Note that in order to find minimal warping error, the image L^* is warped into L that is as similar to segmentation T as possible. The approach to warping a labeling L^* to another labeling L given T is detailed in Algorithm 1, and *simple(L)* indicates the simple points of L. Figure 1 gives a detailed explanation of how the warping error is identified. First, the ground truth annotation in Figure 1(b) is warped into another labeling in Figure 1(c) by using the warping algorithm in Algorithm 1. Then, the topological disagreements presented in Figure 1(e) can be calculated from the pixel differences between the labeling in Figure 1(c) and the labeling in Figure 1(d), which contains two topological errors, a merge (blue circle) and a split (red circle). These errors result from the problem of boundary ambiguity in the original image, as can be seen in Figure 1(a).

3 Ground Truth Estimation by Minimizing Warping Error

This section focuses on the main contribution of the paper: estimating a segmentation that is topologically equivalent and geometrically similar to the true, unknown segmentation when a few segmentations are available.

3.1 Problem Definition

Given a set of N segmentations, $S_1^*, ..., S_N^*$, either obtained by automated segmentation algorithms or annotated by different humans, the goal is to find an estimated segmentation $\hat{S}$ that is topologically equivalent and geometrically similar to the underlying unknown true segmentation. One potential segmentation that satisfies the topological and geometrical constraints and is capable of representing the true, unknown segmentation is that with a topology most of the given segmentations agree on. In other words, the estimated segmentation $\hat{S}$ is a segmentation that minimizes the warping error between itself and the given segmentations. Mathematically, $\hat{S}$ is obtained by minimizing the following:

$$\hat{S} = \underset{S}{\operatorname{argmin}} \sum_{i=1}^{N} D(S \parallel S_i^*) = \underset{S}{\operatorname{argmin}} \sum_{i=1}^{N} \min_{S_i \triangleleft S_i^*} |E(S, S_i)|, \tag{2}$$

where S_i is the optimal warping of the labeling S_i^*.

One possible method to find the estimated segmentation $\hat{S}$ is to enumerate all possible labelings and choose one that has minimal warping error. However, enumerating all labelings can be computationally expensive. Another alternative is to gradually change the topology of a segmentation and make it converge to a topology that most segmentations agree on. The section below details this approach.

3.2 Proposed Topological Correction Algorithm

Changing the labeling of an image involves flipping the labels of pixels, which can result in a merger of two adjacent regions or a splitting of a region into two. The potential pixels whose change of label causes a topological change are those that affect warping error. To achieve the goal of seeking an estimated segmentation with a topology that most segmentations agree on, the algorithm starts with an initial segmentation obtained by using the majority voting method. At each iteration, by using the number of topological errors as the evaluation metric, the algorithm corrects one topological disagreement between the estimated segmentation and the given segmentations. While correcting a topological disagreement at each iteration, the algorithm selects an error having a lowest flipping cost defined in Equation 4, detailed in the next section. A new labeling is accepted only if it has less warping error. The algorithm repeats the process of correcting topological errors and stops when no topological changes can lead to the reduction in the overall warping error, that is, it reaches a segmentation that minimizes warping error defined in Equation 2. Algorithm 2 details this proposed method.

Algorithm 2. Topological correction by minimizing warping error (proposed algorithm)

input : A set of labled binary images, $S_1^*, ..., S_N^*$
output: An estimate of the ground truth, $\hat{S}$

initialize $\hat{S}$ to the result of majority voting given $S_1^*, ..., S_N^*$;
foreach S_i^* **do**
  $E_i = D(\hat{S} \parallel S_i^*)$;
$E_{\min} = \sum_i E_i$;
while *not converged* **do**
  assign each topological error the flipping cost based on Equation 4;
  select a topological error with the lowest flipping cost;
  flip the selected pixels in the estimated segmentation $\hat{S}$;
  foreach S_i^* **do**
    $E_i = D(\hat{S} \parallel S_i^*)$;
  $E_{\text{new}} = \sum_i E_i$;
  if $E_{\text{new}} < E_{\min}$ **then**
    accept the new estimated ground truth $\hat{S}$;
  else
    reject the new estimated ground truth and restore $\hat{S}$ back to the previous estimation;

3.3 Topological Change Cost

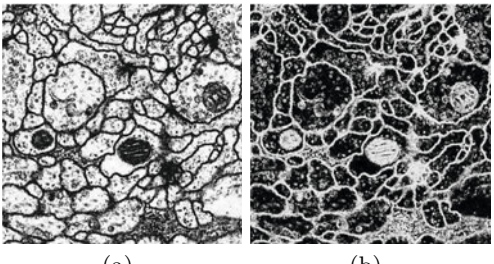
As mentioned above, flipping the label of pixels that contain warping error modifies the topology of a segmentation. To better locate the pixels for topological change, each pixel is associated with a flipping cost, and the selection of what pixels' labels to be changed depends on the cost associated with those pixels. More specifically, the flipping cost of each pixel contains statistical information of an image, such as the intensity distributions.

To define the flipping cost, two notations are first introduced. Let S be the foreground (object) segmentation and $\bar{S}$ be the background segmentation. The cost of flipping the label of a pixel p from S_p to $\bar{S}_p$, $f(p)$, is defined as

$$f(p) = \frac{\Pr(I_p|S)}{\Pr(I_p|\bar{S}) + \Pr(I_p|S)}, \tag{3}$$

where I_p is the intensity value of pixel p, and $\Pr(I_p|S)$ and $\Pr(I_p|\bar{S})$ represents how well the intensity of pixel p fits into the intensity distributions (histograms) of foreground and background, respectively. Because a set of segmentations are given, the intensity histograms for the foreground and background are available. Figure 2(a) shows the flipping cost of changing a label of each pixel from the foreground to the background. Brighter color indicates a higher cost. Similarly, the cost of changing the label of a pixel p from $\bar{S}_p$ to S_p is defined by $\Pr(I_p|\bar{S}) / \left(\Pr(I_p|\bar{S}) + \Pr(I_p|S)\right)$. The flipping cost of changing a label of each pixel from the background to the foreground is shown in Figure 2(b).

As can be seen in the example given in Figure 1, the topological change of a segmentation requires a sequence of pixel flips. Now, let C denote a set of pixels involved in the merger of two adjacent regions or the splitting of a region. To reduce the computational complexity associated with calculating the flipping cost of the pixels, a simple assumption is made that the pixel flip is independent with each other. Therefore, the cost $f(C)$ of flipping all points in C, is defined as sum of the flipping cost of the individual pixels, that is,



(a) (b)

Fig. 2. Flipping cost of changing a label of each point in an image. (a) The flipping cost of changing a label of each pixel from the foreground to the background. (b) The flipping cost of changing a label of each pixel from the background to the foreground. Brighter color indicates a higher cost.

$$f(C) = \sum_{p \in C} f(p) \ . \tag{4}$$

4 Experimental Results

The experiments were carried out on a few synthetic images and on an EM data set [7,8]. The purpose of applying the proposed method to simple synthetic images was to show that the proposed method can retrieve a segmentation that is topologically equivalent to the true, unknown segmentation.

4.1 Synthetic Images

A set of segmentations of an image are required to evaluate the proposed method. Four alternative segmentations were generated, and they are shown in Figures 3(a) to 3(d). The first three segmentations separate the image into two regions (i.e. same topology) with a slightly different boundary localization while the last (Figure 3(d)) segments the image as a whole region (different topology from the rest). These four segmentations present four possible segmentations of an image, and they are to be combined to obtain an estimate of the underlying true segmentation.

Demonstrated in Figure 3 is a simple comparison between the estimated segmentation obtained by the proposed method and those by majority voting and STAPLE. The four alternative segmentations are of high quality, approaching expert levels: The estimated sensitivities (i.e. the probability of an annotator labeling a pixel as foreground if the true label is foreground) were 0.9991, 0.9991, 0.9991, and 1.0000, respectively, and the estimated specificities (i.e. the probability of an annotator labeling a pixel as background if the true label is background) were 1.0000, 1.0000, 1.0000, and 1.0000, respectively. The estimated segmentation of the true, unknown segmentation by using the majority voting method is shown in Figure 3(e), which is a segmentation that most segmentations agree on. However, this estimated segmentation is unable to represent the true, unknown segmentation because they are not topologically equivalent. Different from the majority voting method that treats each segmentation equally, STAPLE weights individual segmentation depending on its estimated performance level. The estimated result by using STAPLE is shown in Figure 3(f). However, we can see that although most of the alternative segmentations are topologically correct, such as those in Figures 3(a) to 3(c), STAPLE is unable to produce a topologically correct estimate of the true, unknown segmentation. This indicates its estimation is sensitive to the boundary localization of given labelings. On the other hand, when the same set of initial segmentations were given, our proposed algorithm produced a topologically correct estimate. The resulting segmentation is shown in Figure 3(g), and its respective close-up in Figure 3(j).

4.2 EM Data Set

A serial section Transmission Electron Microscopy (ssTEM) data set of the Drosophila first instar larva ventral nerve cord (VNC) from Cardona et al. [7,8]

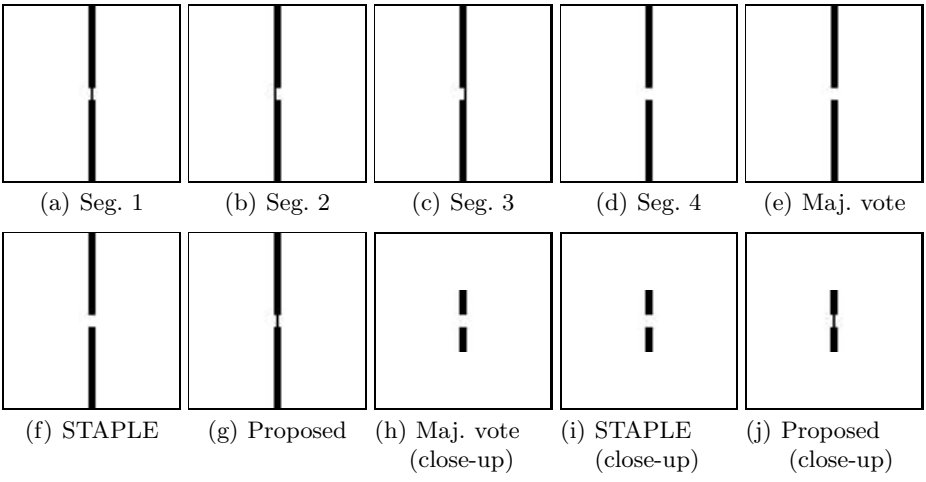
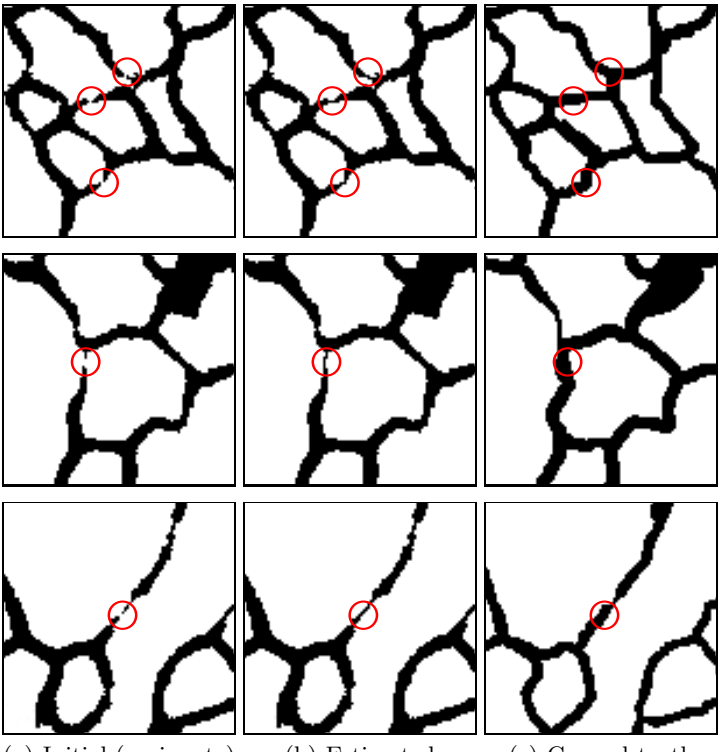


Fig. 3. Comparison between the estimated segmentation obtained by the proposed method and those by majority voting and STAPLE. (a)-(d) Alternative segmentations to be fused. The first three segmentations have the same topology while the last has a different topology from the rest. (e) The estimated segmentation by majority voting merges two separate regions as one, thus causing a merge error. (f) The estimated segmentation by STAPLE also contains a merge error. (g) The estimated segmentation by the proposed method gives a topologically correct estimate. (h)-(j) Close-ups of the results from majority voting, STAPLE, and the proposed algorithm, respectively.

was used for the evaluation of the proposed method. The data set contains 30 sections, each of which having a size of 512×512 pixels. The tissue is $2 \times 2 \times 1.5$ microns in volume, with a resolution of $4 \times 4 \times 50$ nm/voxel. The data set was manually delineated by an expert, and the manual segmentations served as the ground truth the algorithm aims to estimate.

To test the developed method, a number of segmentations were first generated by thresholding the image at different values with additional manual editing to finally construct alternative segmentations. Note that the main focus of our work is not on how these alternative segmentations are generated, so any reasonable manual or automated method will be enough. These generated segmentations represent the alternative segmentations to be combined. Taking those segmentations as input, the proposed method produced an estimated segmentation. Figure 4(a) shows the initial segmentations with which the proposed method starts (majority vote), Figure 4(b) the estimated segmentations obtained by the proposed method, and Figure 4(c) the ground truth annotated by the expert. The topologies of the initial segmentations, obtained by majority voting, do not agree with those of the ground truth, which are indicated by the red circles. The final estimated segmentations, on the other hand, are topologically equivalent and geometrically similar to the ground truth, with minor boundary localization differences. Using the same set of input segmentations and warping error



(a) Initial (maj. vote) (b) Estimated (c) Ground truth

Fig. 4. Comparison of the topologies of initial segmentations, estimated segmentations, and ground truth. (a) The initial estimated segmentations with which the proposed method starts. (b) The estimated segmentations produced by the proposed method. (c) Ground truth. The topologies of the initial segmentations (a) do not agree with those of the ground truth (c), which are indicated by the red circles. The estimated segmentations (b) produced by the proposed method, on the other hand, are topologically equivalent and geometrically similar to the ground truth.

as the evaluation metric, the quantitative comparison of the results obtained by majority voting, STAPLE, and the proposed method is shown in Table 1. As we can see, topological errors exist in the results obtained by majority voting and STAPLE because these two methods fuse segmentation labels at the pixel level. The proposed method, on the contrary, can obtain topologically correct segmentations as long as the topologies of majority of alternative segmentations are correct. Also note that, in the experiment, the proposed method used the majority voting method's results as the initial estimated segmentations and gradually modified the topologies of estimated segmentations until convergence. The proposed method's final estimated segmentations are topologically equivalent to the true segmentations even it started with segmentations containing topological errors.

Table 1. Comparison of the number of topological errors committed by majority voting, STAPLE, and the proposed method on 10 different samples from the EM data set. As we can see, topological errors exist in the results obtained by majority voting and STAPLE whereas the proposed method is able to obtain topologically correct segmentations as long as topologies of most of the alternative segmentations are correct.

Sample #	1	2	3	4	5	6	7	8	9	10
Majority voting	6	4	8	2	2	2	0	2	6	5
STAPLE	2	4	3	5	2	0	3	2	3	2
Proposed method	0	0	0	0	0	0	0	0	0	0

5 Conclusion

We presented a novel pooling method that seeks a segmentation topologically equivalent and geometrical similar to the true, unknown segmentation when a set of alternative segmentations are available. This method is effective for noisy EM images because it maximizes the topological agreements among segmentations during the estimation process and ensures that a segmentation is topologically correct, which is important for connection estimation for connectomics research. Experimental results have demonstrated the effectiveness of this method.

References

1. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23, 903–921 (2004)
2. Artaechevarria, X., Muñoz-Barrutia, A., de Solorzano, C.O.: Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE Trans. Med. Imaging 28, 1266–1277 (2009)
3. Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54, 940–954 (2011)
4. Jain, V., Seung, H.S., Turaga, S.C.: Machines that learn to segment images: a crucial technology for connectomics. Current Opinion in Neurobiology 20, 653–666 (2010)
5. Jain, V., Bollmann, B., Richardson, M., Berger, D.R., Helmstaedter, M.N., Briggman, K.L., Denk, W., Bowden, J.B., Mendenhall, J.M., Abraham, W.C., et al.: Boundary learning by optimization with topological constraints. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2488–2495 (2010)
6. McGuinness, K., O'Connor, N.E.: A comparative evaluation of interactive segmentation algorithms. Pattern Recognition 43, 434–444 (2010)
7. Cardona, A., Saalfeld, S., Tomancak, P., Hartenstein, V.: TrakEM2: open source software for neuronal reconstruction from large serial section microscopy data. In: Proc. High Resolution Circuits Reconstruction, pp. 20–22 (2009)
8. Cardona, A., Saalfeld, S., Preibisch, S., Schmid, B., Cheng, A., Pulokas, J., Tomancak, P., Hartenstein, V.: An integrated micro- and macroarchitectural analysis of the *Drosophila* brain by computer-assisted serial section electron microscopy. PLoS Biol. 8, e1000502 (2010)