

# Segmentation Evaluation for Fluorescence Microscopy Images of Biological Objects

Shantanu Singh<sup>1</sup>, Sundaresan Raman<sup>1</sup>, Jens Rittscher<sup>2</sup>, Raghu Machiraju<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, U.S.A

<sup>2</sup>General Electric Global Research Center, Niskayuna, New York, U.S.A.

## Abstract

Assessing the quality of image segmentation algorithms is an essential step towards the quantitative analysis of biological microscopy images. Given the limited accuracy of segmentation algorithms in all but trivial cases, it is particularly important to define an index to grade the quality of segmentations. Such an index can help guide the choice of algorithms for a particular application, assist in optimizing algorithm parameters, and provide a measure of quality when evaluating scientific conclusions drawn from the results of the segmentation. Motivated by the problem of segmenting microscopy images of thick tissue sections, we propose an approach to evaluate segmentation quality for images that contain a large number of objects (e.g., nuclei). The evaluation of such images is rendered difficult for two reasons, (i) the correspondence of components between two segmentations of the same image is often ambiguous, and (ii) the number of components in the image is typically too large to generate *complete* ground truth for. Existing evaluation techniques of segmentation algorithms are inadequate to be applied under these constraints. Our proposed evaluation strategy addresses both these constraints by suitably modifying a commonly accepted evaluation index. We demonstrate the efficacy of our proposed strategy towards the evaluation of typical segmentations of fluorescence microscopy images of cell nuclei.

Segmentation, Evaluation, Validation, Cell Segmentation, Confocal Microscopy

## I. INTRODUCTION

Image segmentation is a fundamental step in the analysis of biological microscopy images. While there have been several segmentation algorithms proposed in literature, the development of techniques for evaluating segmentation results have received lesser attention until recent years [1], [2]. An objective and quantitative evaluation of segmentations is essential for many reasons. First, it provides an objective criterion for selecting a segmentation algorithm for a specific application. Second, segmentations are the basis of several quantitative biological and clinical studies and it is essential for domain experts to have such an evaluation at hand to judge the validity of these studies. Third, an index can be used to optimize the performance of a segmentation algorithm by guiding the exploration of its parameter space.

The current work is motivated by the problem of segmenting 3D confocal fluorescence microscopy images of cell nuclei. A sample image in Fig. 1(a) shows a region of a mouse mammary tissue section stained with a fluorescent DNA dye. The nuclei segmentations are used in a set of quantitative studies related to tumor progression. Segmentation of cells remains an open image analysis problem that poses several challenges, the crucial ones being (i) dense packing of cells, that make it hard to find a separating boundary, (ii) inhomogeneous staining of cell nuclei, that often results in over-segmentation (iii) irregular cell shapes, that preclude the use of shape priors. These issues are observed in Fig. 1(a). Additionally, in confocal z-stacks, anisotropic sampling and axial signal attenuation further complicate the segmentation problem. Since segmentation algorithms exhibit a degree of imperfection in all but trivial examples, and especially so given the confounds above, performance evaluation becomes a particularly important issue.

Evaluation of segmentation algorithms has been studied in its broader context in pattern analysis and computer vision literature. An early survey of segmentation evaluation methods in pattern analysis literature is provided in [3]. The survey categorizes evaluation techniques into *analytical* methods that evaluate the intrinsic properties of the algorithm, *empirical goodness* methods that evaluate intrinsic properties of the result, and *empirical discrepancy* methods that evaluate results by comparing them with a gold standard, or *ground-truth*. Given the importance of

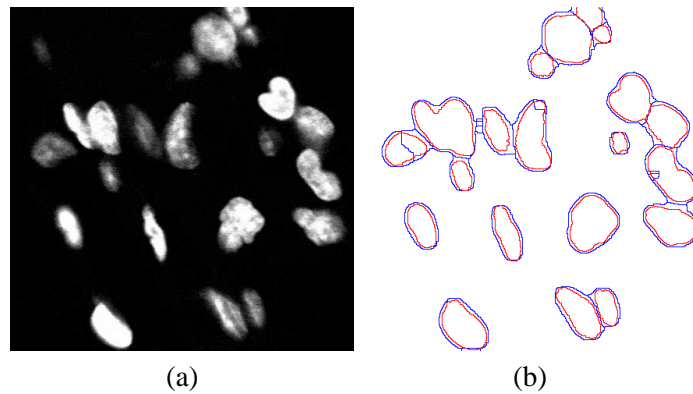


Fig. 1. (a) Confocal fluorescence microscopy image of cell nuclei from a mouse mammary tissue (b) Two segmentations of the image,  $\mathbb{S}$  (red) and  $\mathbb{S}'$  (blue). A component of  $\mathbb{S}$  can overlap with multiple components from  $\mathbb{S}'$  and vice-versa making the correspondence of components between the two segmentations ambiguous. The image corresponds to a  $60\mu\text{m} \times 60\mu\text{m}$  section of a  $146\mu\text{m} \times 146\mu\text{m}$  slice from a stack of 117 slices.

incorporating expert knowledge when evaluating analysis techniques, we focus on discrepancy methods. Therein, the *quality* of a result is measured through its *validity* with respect to a standard. In recent work, Unnikrishnan *et al.* [4] propose a measure of similarity based on the Rand index [4] that allows the comparison of a segmentation with a *set* of ground truth segmentations. Cardoso *et al.* [5] provide a metric that is similar to the classification error distance. Polak *et al.* [6] present a modification of a discrepancy measure proposed in [7] to penalize over-segmentation and under-segmentation.

In medical imaging literature, segmentation evaluation techniques have largely focussed on problems where the number of components in the image are known and limited in number. A typical case is tissue segmentation in a brain magnetic resonance image (MRI) where the label categories are known and fixed (e.g., white matter, gray matter, cerebrospinal fluid, etc.) In these cases, segmentation evaluation is tantamount to measuring *classification* error which can in turn be reported using the canonical measures of sensitivity and specificity. Alternatively, if each class is considered in isolation, indices such as the Dice Similarity Coefficient (DC) or Jaccard Coefficient (JC) [2] can be computed on a per-class basis. Warfield *et al.* [1] propose a framework to estimate a “true” ground truth from a set of ground truth images from different sources, and simultaneously measure the sensitivity and specificity indices of each source. The estimate can thereby be used to evaluate the quality of a segmentation within the same framework [1], or by comparing the segmentation with the estimated ground truth using overlap indices [8], [2], or other similarity measures [8].

The problem of evaluating segmentations in biological microscopy images, specifically those that contain a large number of objects such as fluorescence microscopy images of cell nuclei, poses unique challenges that cannot be addressed directly by the techniques above. First, the number of components in the image, usually in the order of a thousand or higher, make it impractical to collect ground truth for the entire image. For example, the image shown in Fig. 1(a) corresponds to a small section of the complete image stack as indicated. Second, the correspondence between components for a pair of segmentations of the same image is often ambiguous as seen in Fig. 1(b).

The first constraint is often circumvented by basing the evaluation on a smaller, cropped region of the image, for which collecting complete ground truth is feasible. This leads to a restricted evaluation of segmentation quality, and does not characterize the performance of the algorithm on the whole image. Instead, we propose a quality index that can be used as an estimator in the presence of partial ground truth. We empirically show that this estimator is an appropriate choice for our application and can reasonably characterize the quality on the entire image domain through a *uniform random* sampling of the same. The second issue of ambiguous correspondence can be dealt with by appropriately weighing the contribution of all possible correspondences when computing the evaluation index [4], [7], [6] or by establishing an explicit correspondence that is globally optimal in an appropriate sense [5]. We choose the latter approach since the explicit correspondence that it provides can be used to evaluate the segmentation quality locally (on a per object basis), while simultaneously evaluating a global quality measure (explained further in Section II).

The details of the method are provided in the Section II followed by experiments in Section III. We conclude

the paper with Section IV.

## II. METHOD

We denote an image by a set  $X = \{x_i\}_{i=1}^N$  consisting of  $N$  pixels, some of which are foreground and the others background. A *partition* of a set  $X$  is a set of disjoint subsets of  $X$  such that their union is equal to  $X$ . A *foreground segmentation* of  $X$  is a partition of  $X$  containing two sets  $X_F$  and  $X_B$  corresponding to foreground and background pixels respectively. A *complete component segmentation* of  $X$  is a partition of  $X_F$ , each element of which corresponds to an image component. A *partial component segmentation* is defined to be a subset of a complete component segmentation. A *ground-truth component segmentation*  $\mathbb{S}$  is one created by an expert and a *test component segmentation*  $\mathbb{S}'$  is one that is to be evaluated.

Let  $\mathbb{S} = \{S_i\}_{i=1}^K$  and  $\mathbb{S}' = \{S'_j\}_{j=1}^{K'}$  denote a ground-truth component segmentation and test component segmentation respectively. Since the two segmentations are generated based on the same image  $X$ , there should, in the ideal case, be a one-to-one correspondence between the components of  $\mathbb{S}$  and  $\mathbb{S}'$ . However, since in all but trivial cases, the number of components in the two segmentations is different, we instead assume that there is some injective mapping from the elements of  $\mathbb{S}$  to those of  $\mathbb{S}'$ . We seek the injective mapping that maximizes the sum of the volume intersections between corresponding pairs of components. The quality index is then defined as the ratio of this intersection sum to the total volume of all components involved in the domain and range of the optimal mapping. Essentially, we adapt the definition of the Jaccard Coefficient for cases where there are multiple objects in the two segmentations, the number of objects are different, and where the correspondence between objects is not trivially established. The quality index  $R$  of  $\mathbb{S}'$  with respect to  $\mathbb{S}$  is given by

$$R(\mathbb{S}, \mathbb{S}') = \frac{\max_{\pi} \sum_{i=1}^K |S_i \cap S_{\pi(i)}|}{\left| \bigcup_{i=1}^K S_i \cup S_{\pi^*(i)} \right|} \quad (1)$$

where  $\pi : \{1 \dots K\} \rightarrow \{1 \dots K'\}$  is an injective mapping, the maximum is taken over the space of all such mappings, and  $\pi^*$  is the solution to the maximization. The problem of finding  $\pi^*$  can be formulated as a *matching* problem in graphs [9]. Consider a *weighted bipartite graph*  $G$  consisting of two sets of vertices  $V$  and  $V'$ . A vertex  $v$  ( $v'$ ) in  $V$  ( $V'$ ) corresponds to a component  $S$  ( $S'$ ) in  $\mathbb{S}$  ( $\mathbb{S}'$ ). An edge exists between all pairs of  $v$  and  $v'$  with weight  $|S \cap S'|$ . The optimization problem of finding  $R$  is equivalent to finding a set of edges such that no two edges are incident on the same vertex and the sum of the weights of chosen edges is maximized. Such a set of edges is termed as a *maximum matching* of  $G$ . This is an instance of the assignment problem [9] which can be solved in  $O((K + K')^3)$  time using a combinatorial optimization algorithm [9]. Fig. 2 shows the bipartite graph for the pair of segmentations in Fig. 1(b).

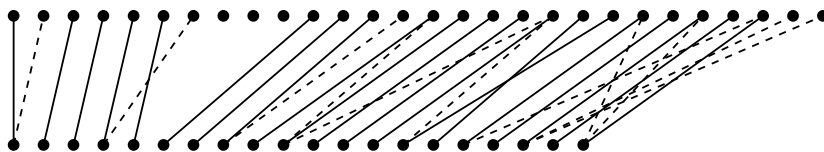


Fig. 2. Bipartite graph corresponding to Fig. 1(b). The bottom (top) row of vertices corresponds to the components of  $\mathbb{S}$  ( $\mathbb{S}'$ ). Edge weights have been left out for clarity. Edges with zero weight have not been drawn. The dark edges correspond to the *maximal matching* solution.

We next consider the case when the quality is assessed on a sparse sampling of the set of components in the image. Let  $\mathbb{S}_{full}$  and  $\mathbb{S}_{part}$  denote ground truth that are complete and partial component segmentations respectively. Let  $R_{full}$  and  $R_{part}$  denote the quality of  $\mathbb{S}'$  with respect to  $\mathbb{S}_{full}$  and  $\mathbb{S}_{part}$  respectively.  $\mathbb{S}_{full}$  contains all the components in the image and hence is a (finite) *population* of our study.  $\mathbb{S}_{part}$  corresponds to a (small) *sample* derived from the population.  $R_{full}$  is a population parameter and  $R_{part}$  is a sample statistic used to estimate  $R_{full}$ . We note the following caveat with using this estimator. Let  $M_{part}$  and  $M_{full}$  correspond to the maximum matching of  $\mathbb{S}'$  with  $\mathbb{S}_{part}$  and  $\mathbb{S}_{full}$  respectively.  $R_{part}$  and  $R_{full}$  are essentially the average of the weights of the

edges in  $M_{part}$  and  $M_{full}$  respectively. We note that  $S_{part} \subset S_{full} \not\Rightarrow M_{part} \subset M_{full}$ . Thus, while the estimator is consistent (the estimate converges to the parameter), its properties of efficiency and bias cannot be trivially analyzed. However, through an empirical study of the sampling distribution of  $R_{part}$ , we have observed that its variance is within acceptable bounds for a reasonable sample size (see Section III). Hence, we argue that  $R_{part}$  is empirically a good choice for an estimator of  $R_{full}$ .

### III. EXPERIMENTS

The problem of evaluating segmentations of biological microscopy images has been motivated by a cellular morphology study related to tumor progression and is based on images from tissue sections of mouse mammary glands using confocal fluorescence microscopy. The tissues are stained using a DNA-specific fluorescent dye, DRAQ5 (Biostatus, Shepshed, UK) and  $z$ -sections have been imaged using 63x/1.3 NA oil objective with a Zeiss 510 Meta confocal microscope at an in-plane resolution of  $0.14\mu m \times 0.14\mu m$  and between-plane resolution of  $0.33\mu m$ .

We first evaluate segmentation results of two algorithms on a sample image. Algorithm 1 uses local adaptive thresholding followed by morphological erosion and median filtering. Algorithm 2 uses histogram equalization followed by Otsu thresholding. Fig. 3 shows the resulting segmentation masks on section of a single slice of a confocal  $z$ -stack. Based on the proposed index  $R$ , Fig. 3(a) rates as the best segmentation result, which agrees with a subjective notion of segmentation quality. For example, note that (a) and (e) have very similar segmentation masks, except that (a) is able to separate the two closely touching nuclei in the top left of the image, whereas (e) does not.

Next, we empirically study the properties of the estimator proposed in Section II. We consider  $R_{part} = R(S_{part}, S')$  as a *sample statistic* of the *sample*  $S_{part}$  and empirically compute its sampling distribution. To do so, a partial ground truth segmentation was created by an expert by marking segmentation masks for an arbitrary set of cell nuclei in an image. By an independent random sampling of this set, several samples of a fixed size were generated. Each of these samples constitute an instance of a partial ground-truth  $S_{part}$ . For a given test segmentation, the  $R_{part}$  was computed by comparing with each instance of  $S_{part}$ . The mean and standard deviation of the  $R_{part}$  values were computed. We do this across three sample sizes and three test segmentations. The results are summarized in Table I. As the values indicate, the standard error of the statistic is within acceptable limits for the given sample size.

	20	30	40
$S'_1$	$0.78 \pm 0.03$	$0.82 \pm 0.02$	$0.81 \pm 0.02$
$S'_2$	$0.55 \pm 0.03$	$0.54 \pm 0.02$	$0.55 \pm 0.03$
$S'_3$	$0.75 \pm 0.04$	$0.73 \pm 0.03$	$0.76 \pm 0.02$

TABLE I  
SUMMARY OF THE SAMPLING DISTRIBUTION OF  $R_{part}$ .

### IV. DISCUSSION AND CONCLUSION

The paper presents an approach to evaluate segmentation results of biological microscopy images. The approach focuses on cases where the number of components in the image is too large to generate complete ground truth for and the correspondence between components of two segmentations is unknown. Existing techniques cannot be adequately employed under these constraints. A similarity index (also used as a quality index) is proposed that finds a matching between components of the test segmentation and ground truth segmentation by maximizing the average component-wise similarity of matched pairs. Importantly, the quality index can also be estimated in the presence of *partial* ground truth. Empirical results indicate that the variance of the estimate remains within acceptable bounds. The segmentation evaluation approach generalizes to images that exhibit the characteristics indicated above.

### REFERENCES

- [1] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 7, pp. 903–921, 2004.

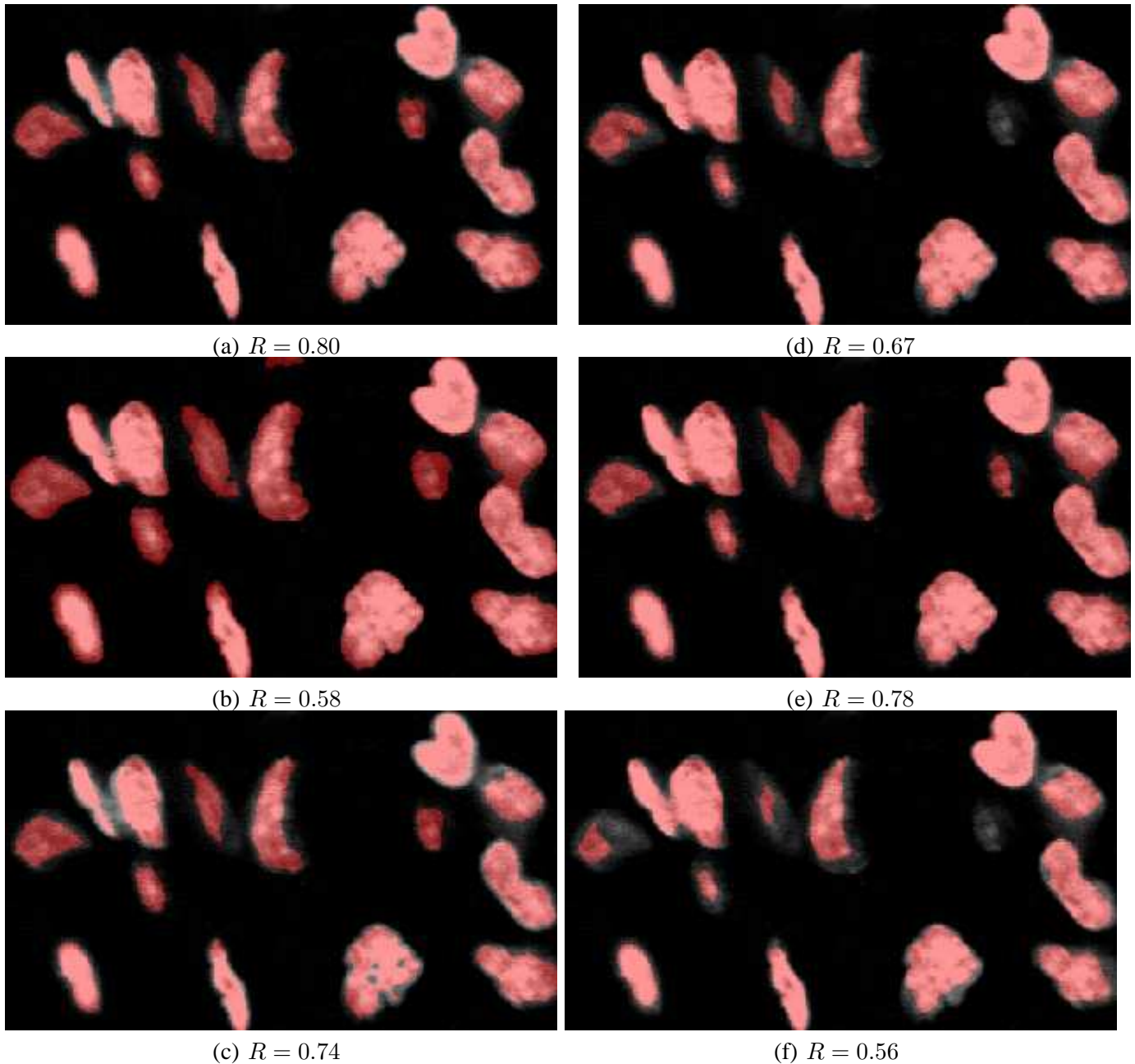


Fig. 3. Segmentation evaluation on cell nuclei images. Segmentations (a)–(c) and (d)–(f) were generated using Algorithm 1 and Algorithm 2 respectively. The quality index of  $R$  of each of the segmentations is indicated.

- [2] S. Bouix, M. Martin-Fernandez, L. Ungar, M. Nakamura, M. S. Koo, R. W. McCarley, and M. E. Shenton, “On evaluating brain tissue classifiers without a ground truth,” *Neuroimage*, vol. 36, no. 4, pp. 1207–1224, July 2007.
- [3] Y. Zhang, “A survey on evaluation methods for image segmentation,” *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, August 1996.
- [4] Ranjith Unnikrishnan, Caroline Pantofaru, and Martial Hebert, “Toward objective evaluation of image segmentation algorithms,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 929–944, 2007.
- [5] J. S. Cardoso and L. Corte-Real, “Toward a generic evaluation of image segmentation,” *Image Processing, IEEE Transactions on*, vol. 14, no. 11, pp. 1773–1782, 2005.
- [6] M. Polak, H. Zhang, and M. Pi, “An evaluation metric for image segmentation of multiple objects,” *Image and Vision Computing*, October 2008.
- [7] David R. Martin, *An Empirical Approach to Grouping and Segmentation*, Ph.D. thesis, EECS Department, University of California, Berkeley, Aug 2003.
- [8] Kelly H. Zou, William, Ron Kikinis, and Simon K. Warfield, “Three validation metrics for automated probabilistic image segmentation of brain tumours,” *Statistics in Medicine*, vol. 23, no. 8, pp. 1259–1282, 2004.
- [9] Douglas B. West, *Introduction to Graph Theory (2nd Edition)*, Prentice Hall, August 2000.