NUCLEI SEGMENTATION IN MICROSCOPE CELL IMAGES: A HAND-SEGMENTED DATASET AND COMPARISON OF ALGORITHMS

Luís Pedro Coelho[1,2,3], *Aabid Shariff*[1,2,3], *Robert F. Murphy*[1,2,3,4]

[1]Lane Center for Computational Biology, Carnegie Mellon University
[2]Center for Bioimage Informatics, Carnegie Mellon University
[3]Joint Carnegie Mellon University–University of Pittsburgh PhD. Program in Computational Biology
[4]Depts. of Biological Sciences, Biomedical Engineering, and Machine Learning at Carnegie Mellon University

ABSTRACT

Image segmentation is an essential step in many image analysis pipelines and many algorithms have been proposed to solve this problem. However, they are often evaluated subjectively or based on a small number of examples. To fill this gap, we hand-segmented a set of 97 fluorescence microscopy images (a total of 4009 cells) and objectively evaluated some previously proposed segmentation algorithms.

We focus on algorithms appropriate for high-throughput settings, where only minimal user intervention is feasible.

The hand-labeled dataset and all the software used to compare different will be made available online at the time of publication. This will enable others to use our dataset as a benchmark against for newly proposed algorithms.

***Index Terms*—** Biomedical image processing, Image segmentation

1. INTRODUCTION

Nuclear level segmentation is an important step in the pipeline of many analytical cytology analyses. It forms the basis of many simple operations (cell counting, cell-cycle assignment,...) and, often, the first step in cell-level segmentation. Proposed algorithms are too often evaluated subjectively or based on a few examples. In order to objectively evaluate segmentation algorithms, we built a dataset of hand-segmented fluorescence microscopy images.

We also evaluated some published algorithms for this problem against our hand-labeled dataset. We were interested in algorithms that were applicable to large-scale automated data collection. Therefore, while parameter tuning for the properties of a given image collection was an acceptable burden on the human operator, tuning for single images was not.

Related Work

Pascal Bamford undertook a similar effort in bright-field microscopy images of cell nuclei [1]. Recently, Gelasca et al. made a available a series of ground truth assignments for different tasks in bioimage segmentation [2], but it did not include a dataset of hand-labeled single nuclei. Our dataset is thus a complement to their work.

	U20S	3T3
Pixel size	1349×1030	1344×1024
Nr. Cells	1831	2178
Avg. Cover	23%	18%
Min Nr. Cells	24	29
Max Nr. Cells	63	70

Table 1. Main Properties of the Two Collections. Avg. cover denotes the percentage of pixels covered by cells. The minimum and maximum are over all the images in each collection.

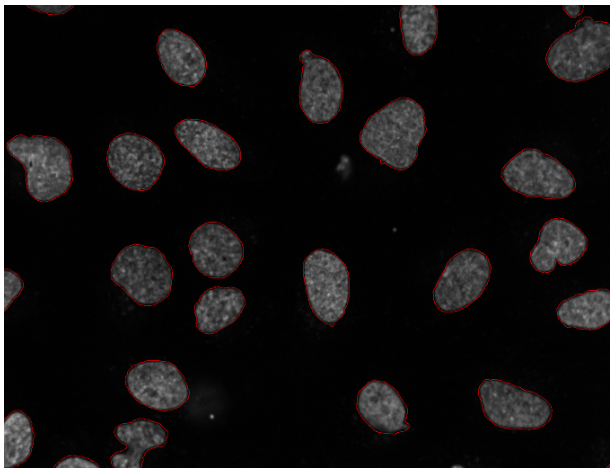
2. DATASET

The dataset is composed of two different collections. The first collection is of U20S cells, previously described in the work by Peng et al [3]. Figure 1 shows two images from this collection. An initial set of 50 images from this collection was chosen, but 2 images were rejected as containing no in-focus cells. Table 1 lists the main properties of the two collections.

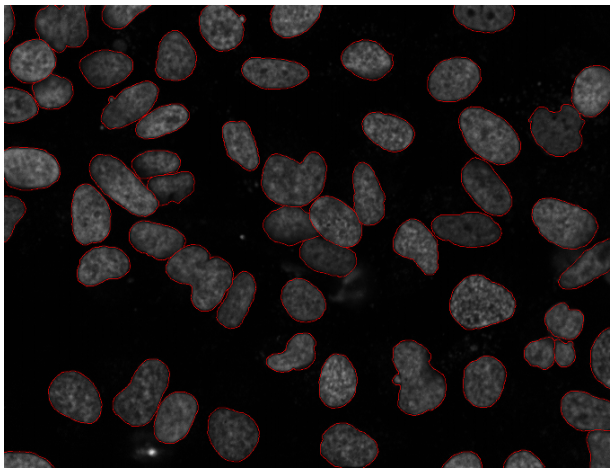
.

The second collection is cell nuclei from NIH3T3 cells, collected using the methodology reported by Osuna et al. [4]. Nuclei in this group are further apart and there is less clustering. They are also more homogeneous in shape and size (not shown). On the other hand, nuclei in single images vary greatly in brightness and images often contain visible debris. Therefore, we consider this a more challenging dataset for automated methods. As with the U20S collection, 50 images were initially chosen, but 3 were rejected as containing no in-focus cells.

In all cases, manual segmentation was performed without access to the protein channel. All images were segmented by LPC. Independently, AS segmented 10 images (5 from each



(a) "Easy" image



(b) "Difficult" image

Fig. 1. Two example images from the U20S collection. (a) shows nuclei that are well separated. Automatic segmentation is expected to do well. (b) has many clustered nuclei and is expected to challenge segmentation algorithms. Most images in the collection lie in between these two examples.

collection).

3. METHODS

3.1. Segmentation

Due to space limitations, we will only describe the ways in which our implementation was adapted to our images and refer the reader to the original publications for detail.

3.1.1. Thresholding

As a baseline algorithm, we implemented a thresholding-based algorithm, which computes a threshold using one of 3 methods: Ridler-Calvard [5], Otsu [6], or the mean pixel value. All above-threshold contiguous regions are considered objects. To remove some noise, we filter the thresholded image with a median filter (window of size 4) and remove all small objects. The size threshold was set to 2500 pixels, circa 64 square microns.

3.1.2. Seeded watershed

We implemented two versions of seeded watershed, both run on a thresholded version of the image (using the mean as threshold, which, as we show below, is a better thresholding method for these images). One operates directly on a blurred version of the image[1], while the second one operates on the gradient of the image. In both cases, seeds are regional maxima of the blurred image. As above, objects too small to be nuclei were removed.

3.1.3. Active masks

Active masks are a recent proposal by Srinivasa et al. [7]. The algorithm assumes that there are two classes of objects, foreground and background. Its only parameters are the mean value and standard deviation of the background region.[2]

Manual tuning led to the following semi-automatic procedure for parameter setting: the value of the background mean is assumed to be the histogram peak plus 3, while the background standard deviation is set to 0.5.

3.1.4. Lin's Merging Algorithm

Lin et al. published an algorithm that is based on merging multiple regions obtained from watershed, using shape information learnt from a dataset [8]. We have implemented a slight variation of their algorithm, but retained the structure. In particular, we use mean method for segmentation, and as shape features: fraction of area that is contained in the convex hull, roundness, eccentricity, area, perimeter, semi-major, and

[1] We used a gaussian blur with a width of 12 pixels.

[2] The active mask framework is more general than this, but we restrict ourselves to the original proposal.

semi-minor axes (all, except the first, computed on the convex hull). Apart from these minor changes, the algorithm is unchanged.

For evaluation, we trained the system using the images segmented by AS and tested on the images segmented by LPC.

3.2. Evaluation

Several metrics have been proposed for evaluation of segmentation results against a hand-labeled gold standard. Some approaches stem from viewing segmentation as a form of clustering of pixels. This allows the use of metrics developed for the evaluation of clustering results. From this family of approaches, we used the Rand and Jaccard indices [9, 10].

The disadvantage of such metrics is that they do not take into account the spatial characteristics of segmentation. In fact, the exact location of the border between foreground and background is often fuzzy. An algorithm that returns a nucleus which almost matches the gold standard except for a one-pixel-wide sliver around the border should be judged very highly even if that sliver contains a large number of pixels. Pascal Bamford, whose work on evaluation of bright-field microscopy is similar to our work on fluorescence microscopy, used the Hausdorff metric [1].

3.2.1. The Rand and Jaccard Indices

Let S be the segmented image and R be the reference image. Let (i, j) range over all pairs of pixels where $i \neq j$, then each pair falls into one of four categories: (a) $R_i = R_j$ and $S_i = S_j$, (b) $R_i \neq R_j$ and $S_i = S_j$, (c) $R_i = R_j$ and $S_i \neq S_j$, (d) $R_i \neq R_j$ and $S_i \neq S_j$. If we let a, b, c, d refer to the number of pairs in its corresponding category, then the Rand index is defined as:

$$\mathrm{RI}(R, S) = \frac{a + d}{a + b + c + d}. \tag{1}$$

That is, the Rand index measures the fraction of the pairs where the two clusterings agree. The Rand index ranges from 0 to 1, with 1 corresponding to perfect agreement.

Based on the same definitions for a, b, c, d, the Jaccard index is defined as:

$$\mathrm{JI}(R, S) = \frac{a}{a + b + c}. \tag{2}$$

The Jaccard index is not upper

3.2.2. Error Counting

Each object in the segmented image is assigned to the object in the reference image with which is share the most pixels. Based on these assignments, we can define the following classes of errors: **split** two segmented nuclei are assigned to a single reference nucleus; **merged** two reference nuclei are assigned to a single segmented nucleus; **add** a segmented nucleus is assigned to the reference background; **missing nucleus** a reference cell is assigned to the segmented background.

3.2.3. Spatially-Aware Evaluation Methods

We implemented two spatially-aware evaluation metrics. Both are based on assigning segmented nuclei to reference nuclei as above, as they are computed between pairs of matched objects. The Hausdorff metric is computed as described by Bamford [1].

Similarly, for each unmatched pixel, we compute its distance to the border. The normalised sum of distances is then defined as:

$$NSD(R, S) = \frac{\sum_i [\![R_i \neq S_i]\!] * D(i)}{\sum_i D(i)}, \tag{3}$$

where the sum index i ranges over pixels and $D(i)$ is the distance of pixel i to the border of the reference object. From the equation, it is obvious that $ND(R, S) \in [0, 1]$, with 1 corresponding to perfect agreement and 0 to no-overlap.

4. RESULTS

Table 2 summarises the results obtained.

Both manual segmentations are in general agreement. Disagreements can be tracked down to an image where the authors differed on whether some small bright objects should be marked as nuclei or debris.

Both Otsu and Ridler-Calvard thresholding score poorly, missing many cells, particularly in the 3T3 collection. In this collection, the presence of very bright cells leads the algorithm to set a threshold between the very bright cells and the rest of the cells, instead of setting it between the foreground and background. The mean thresholding is better suited for these images, which consist mainly of background with objects of very different intensities.

Watershed does poorly and does not justify the cost over the simpler mean-based segmentation. Active masks score poorly mainly due to nuclei over-segmentation. In the 3T3 collection, it also adds many spurious elements. Finally, Lin's merging algorithm obtains very good results in the U20S case, but the debris in the 3T3 collection leads to main spurious objects being tagged as nuclei.

We also notice the Rand and Jaccard indices while distinguishing the alternative manual segmentation from the automatic ones are not good measures for this data as they fail to distinguish between the better and the worse algorithms. Both the Hausdorff and the NSD metrics capture the relationships between the algorithms well.

Algorithm	RI	JI	Hausdorff	NSD	Split	Merged	Spurious	Missing
AS manual threshold	94%/95%	3.4/2.4	12/10	0.1/0.0	1.4/2.8	1.2/1.0	1.0/2.4	2.8/2.8
RC threshold	73%/89%	2.1/2.1	39/43	0.4/0.2	0.7/1.0	1.7/2.1	1.5/0.3	29.7/8.3
Otsu threshold	75%/92%	2.1/2.2	37/35	0.3/0.1	0.8/1.1	2.1/2.4	1.7/0.3	26.6/5.6
Mean threshold	82%/96%	1.9/2.2	24/27	0.2/0.1	1.5/1.3	5.1/3.4	3.1/0.9	4.8/3.7
Watershed (direct)	80%/95%	1.8/2.2	27/38	0.5/0.6	2.8/12.8	3.0/1.4	24.8/25.4	0.0/0.0
Watershed (gradient)	81%/95%	1.8/2.2	30/33	0.5/0.6	2.7/7.6	3.6/2.0	25.8/31.8	0.0/0.0
Active Masks	77%/93%	1.6/2.0	118/146	0.7/0.6	10.7/18.8	4.6/3.9	33.6/6.3	1.2/0.2
Lin's Merging	71%/88%	2.0/2.1	34/21	0.6/0.6	1.5/2.9	1.1/1.2	8.1/26.7	4.6/0.0

Table 2. Comparison of Segmentation Algorithms. For the algorithms considered in this paper, this presents the result of segmenting compared against the hand-segmented gold-standard. In each table cell, we input two values corresponding to the two datasets used, 3T3 and U20S.

5. DISCUSSION

We presented a dataset that can be used to evaluate cell nuclei segmentation algorithms. This dataset consists of two collections, from different cell types and different microscopes.

We also implemented several published algorithms for cell nuclei segmentation and tested them against our gold-standard. No single algorithm outperformed all others in both datasets.

The hand-labeled dataset will be made available online at the time of publication. We will also make available all the software necessary to generate all the computational results in this paper.

5.1. Acknowledgements

LPC was partially funded by the Fundação Para a Ciência e Tecnologia (grant SFRH/BD/37535/2007) as well as a fellowship from the Fulbright Program.

6. REFERENCES

[1] P. Bamford, "Empirical comparison of cell segmentation algorithms using an annotated dataset," in *Image Processing, 2003. ICIP 2003. Proc. 2003 International Conference on*, 2003, vol. 2, pp. II–1073–6 vol.3.

[2] Elisa Drelie Gelasca, Jiyun Byun, Boguslaw Obara, and B.S. Manjunath, "Evaluation and benchmark for biological image segmentation," in *IEEE International Conference on Image Processing*, Oct 2008.

[3] Tao Peng, Ghislain M.C. Bonamy, Estelle Glory, Sumit K. Chanda Daniel Rines, and Robert F. Murphy, "Automated unmixing of subcellular patterns: Determining the distribution of probes between different subcellular locations," *Proc. of the National Academy of Sciences (Submitted)*, 2009.

[4] Elvira García Osuna, Juchang Hua, Nicholas Bateman, Ting Zhao, Peter Berget, and Robert Murphy, "Large-scale automated analysis of location patterns in randomly tagged 3t3 cells," *Annals of Biomedical Engineering*, vol. 35, no. 6, pp. 1081–1087, Jun 2007.

[5] T. Ridler and S. Calvard, "Picture thresholding using an iterative selection method," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 8, no. 8, pp. 630–632, Aug. 1978.

[6] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, January 1979.

[7] Gowri Srinivasa, Matthew C. Fickus, Manuel N. Gonzalez-Rivero, Sarah Yichia Hsieh, Yusong Guo, Adam D. Linstedt, and Jelena Kovacevic, "Active mask segmentation for the cell-volume computation and golgi-body segmentation of hela cell images," in *ISBI*. 2008, pp. 348–351, IEEE.

[8] Gang Lin, Umesh Adiga, Kathy Olson, John F. Guzowski, Carol A. Barnes, and Badrinath Roysam, "A hybrid 3d watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks," *Cytometry Part A*, vol. 56A, no. 1, pp. 23–36, 2003.

[9] William M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, Dec., 1971.

[10] Gilbert Saporta and Genane Youness, "Comparing two partitions: Some proposals and experiments," in *Proc. in Computational Statistics*. 2002, Physica Verlag.