Predicting Segmentation Accuracy for Biological Cell Images

Adele P. Peskin[1], Alden A. Dima[2], Joe Chalfoun[2], and John T. Elliot[2]

[1] NIST, Boulder, CO 80305
[2] NIST, Gaithersburg, MD 20899

Abstract. We have performed segmentation procedures on a large number of images from two mammalian cell lines that were seeded at low density, in order to study trends in the segmentation results and make predictions about cellular features that affect segmentation accuracy. By comparing segmentation results from approximately 40000 cells, we find a linear relationship between the highest segmentation accuracy seen for a given cell and the fraction of pixels in the neighborhood of the edge of that cell. This fraction of pixels is at greatest risk for error when cells are segmented. We call the ratio of the size of this pixel fraction to the size of the cell the extended edge neighborhood and this metric can predict segmentation accuracy of any isolated cell.

1 Introduction

Cell microscopy is being used extensively to monitor cellular behavior under experimental settings. The common use of CCD cameras and the availability of microscopes with excellent optics, light sources, and automated stages and filter wheels allows collection of quantitative multiparameter image sets of large numbers of cells [1]. When combined with image analysis procedures, these image sets can provide several measurements of cellular behavior under the experimental conditions.

One of the most common image analysis procedures for cellular images is segmentation of an image object from the remainder of the image. For example, for images of cells stained with a fluorescent dye that covalently attaches to cellular proteins [2], segmentation procedures can be used to identify image pixels that are associated with the cell and separate them from the background. The results of this type of segmentation on images of cells that are at low density (i.e. minimal cell-cell contact) can be used to generate metrics such as spreading area, cell shape, and edge perimeter and provide high quality information about the morphological shape of the cells. This information is characteristic of cell phenotype or state and provides measurements that can be used to compare experimental conditions [3].

There is extensive literature on procedures used to segment whole cells. Depending on the properties of the imaged cells, automated segmentation algorithms can provide differential results even when applied to the same image.

New sophisticated algorithms such as level sets and active contours can segment particular cell features, but they are not necessarily readily available for conventional users of image analysis software. Automated segmentation routines based on histogram analysis or simple gradient-based edge detection routines are more common in most image analysis software. Although these methods can provide appropriate cell segmentation under many conditions, these segmentation routines often fail to adequately segment certain cells. In this study, we segmented a large number of cells from cellular images with various algorithms to identify what features in a cell object can influence segmentation outcome for a particular cell. A metric that evaluates the staining properties at the edge of a cell was developed to score the cell edge properties. This metric, called the edge neighborhood fraction, was found to be predictive of segmentation accuracy under several experimental conditions.

Because image segmentation is critical to biological image analysis, many segmentation methods have been published, including histogram-based, edge-detection-based, watershed, morphological, and stochastic techniques [4]. There are few examples of systematic comparisons of image analysis algorithms for cell image data; these include a comparison of cell shape analysis, a recent report comparing segmentation algorithms [5] and a study on predicting confidence limits in segmentation of cells [6]. An ongoing study in our group compares nine different segmentation techniques to manually segmented cells on a small number of cell images [7]. The study presented here evaluates cells from images of two different cell lines under five different sets of imaging conditions. Overall, we evaluated over 40000 cells from both NIH3T3 fibroblast and A10 smooth muscle cells in 9000 images. The 40000 cells represent 4 replicate wells of each of 2000 unique cells imaged under five different settings which varied edge quality. A study on this scale was large enough to produce statistically reliable results about the accuracy of the segmentation methods evaluated here and form predictive information for individual cells in different cell lines over a range of representative imaging conditions.

2 Data description

The data used in this study examine two cell lines and five imaging conditions. These images consist of A10 rat smooth vascular muscle cells and NIH3T3 mouse fibroblasts stained with a fluorescent Texas Red-C2-maleimide cell body stain [2]. The overall geometric shape of the cell lines differ. A10 cells are well spread large cells. NIH3T3 cells are smaller fibroblasts with a spindly shape. The five imaging conditions varied in both the exposure time and the filter settings. These settings resulted in varying the line per mm resolution and the signal to noise ratio. Multiple cells are present on most images.

Both cell lines were maintained as previously described [8]. Cells were seeded at 1200 (NIH3T3) or 800 (A10) cells/cm^2, in 6 well tissue culture PS plates and incubated overnight. The cells were fixed with 1% PFA in PBS, and stained with Texas-Red maleimide and DAPI as previously described [2]. Images of the

stained cells were collected with an Zeiss automated microscope with automated filter wheels controlled by Axiovision software. For optimal filter conditions, the stained cells were visualized with a Texas Red filter set (Chroma Technology, Excitation 555/28, #32295; dichroic beamsplitter #84000; Emission 630/60, #41834). For non-optimal filter conditions, the cells were imaged with Texas Red excitation filter (Chroma Technology, Excitation 555/28 filter dichroic beamsplitter #84000; Emission 740lp, #42345). These imaging conditions result in reduced intensity signal to noise ratios and introduce blurring.[3] Exposure times were selected to use either 1/4, full, or partially saturated dynamic range of the CoolSnap HQ2 camera. The five imaging conditions are summarized in Table 1.

To segment each of the cells and generate reference data that closely mimics human drawn segmentations, we used a procedure that was developed based on the analysis of manual segmentation processes. The procedure is described in reference [9]. This algorithm was applied to images with the highest contrast (i.e. conditions 3, Table 1) and edge pixels were identified as pixels with at least one neighbor pixel with an intensity less than 0.7 of its adjacent value. This intensity gradient feature has been shown to correlate well with manually selected edge pixels. Figure 1 shows a typical image and the reference data cell masks.

Table 1. The five sets of imaging conditions.

Image	Exposure time(s) A10	Exposure time(s) NIH3T3	Filter type
1	0.015	0.01	optimal filter
2	0.08	0.05	optimal filter
3	0.3	0.15	optimal filter
4	1.0	1.0	non-optimal filter
5	5.0	5.0	non-optimal filter

3 Extended Edge Neighborhood

The apparent cell edges vary widely in clarity and sharpness across the five different images of the same cells. In particular, the images vary in terms of the number of pixel lengths (distance between pixels) needed to represent the thickness of the edge regions of the cells. We have previously quantified this thickness with a metric we call the cell edge quality index (QI) [10]. In the next section we describe how we find and quantify the fraction of pixels that are at risk for inaccuracy during a segmentation, using the quality index and the cell geometry.

[3] Certain trade names are identified in this report only in order to specify the experimental conditions used in obtaining the reported data. Mention of these products in no way constitutes endorsement of them. Other manufacturers may have products of equal or superior specifications.

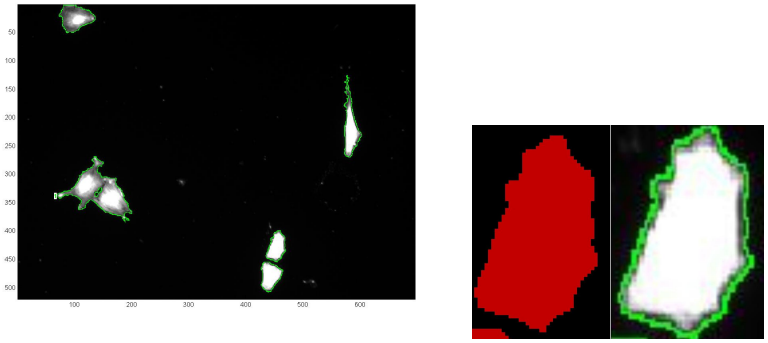

Fig. 1. The outlines in green of masks resulting from our semi-automated method for mimicing the manual segmentation process; A close-up of one cell near the bottom of the image, with the mask in red and the cell outline in green.

In an accompanying paper, we compare manual segmentation masks on a set of 16 of the images used here, with type 3 imaging conditions from Table 1. We find that the our manual segmentation sets differ from one another, and that the extent of misalignment from one set to another depends upon the cell geometry [7] [9]. We observed that smaller, less round cells were more at risk for error in the selection. The smaller the cell, the less likely that two people hand selecting a mask would pick a large fraction of exactly the same pixels for the mask: if most of the pixels are near an edge, a cell is more at risk for any kind of segmentation error. In addition to cell size and shape, the gradient of the pixel intensity at the cell edge also plays a large role in determining whether a cell image can be segmented properly. We combine these concepts into a single quantity that can be calculated quickly for each cell in an image. The metric represents the size of the extended edge neighborhood of pixels and is a fraction derived from the ratio of pixels at risk to the total area of the cell. The pixels at risk are determined by multiplying the cell perimeter by a factor determined from the quality index (QI), that represents the physical thickness (Th) of the edge of the cell.

4 Quality Index and Edge Thickness Calculation

For each cell in an image, we evaluate the pixel intensities within an isolated region containing the cell and background pixels. The quality index is calculated as follows [10]:

1. Identify the 3-component Gaussian mixture, whose components correspond to background (B), edge (E), and cell (C) pixels, via the EM (Expectation-Maximization) algorithm; x_B, x_E, and x_C denote the means of each component [11], [12].
2. Find the average gradient magnitude at each intensity between x_B and x_E.

3. Smooth the gradient in this region to fill in any gaps, and denote the resulting function by G(Intensity).
4. Find the intensity, Intensity A, at which the smoothed gradient magnitude is maximized.
5. Find the expected neighboring pixel to a pixel with Intensity A and denote this intensity as B; i.e., Intensity B = A - G(A)*(1 pixel unit).
6. Find the expected neighboring pixel to a pixel with Intensity B; i.e., Intensity C = B - G(B)*(1 pixel unit) = A - G(A)*(1 pixel unit) - G(A-G(A)*(1 pixel unit))*(1 pixel unit).
7. Compute the quality index as QI = (A - C)/(A - x_B).

The quality index ranges from 0.0 to 2.0, with a perfectly sharp edge at a value of 2.0. The edge thickness is defined as Th = 2.0/QI to scale a perfectly sharp edge to be equal to 1.0 pixel unit. We approximate the number of pixels at the edge by multiplying the edge thickness, Th, by the cell perimeter, and then define our new metric, the ratio of pixels at the edge to the total number of pixels, the extended edge neighborhood (EEN), as:

$$EEN = (P \times Th)/area \tag{1}$$

This value is effectively the fraction of the total cell area that makes up the cell edge. We determine the cell perimeter and area from a 5-means clustering segmentation mask. This is a simple, fast segmentation, which we know from our ongoing study [7] is consistently a good performing algorithm over a wide range of extended edge neighborhoods and provides reliable estimates of cell areas and perimeters values.

The extended edge neighborhood is influenced by the intensity contrast of the image, through the calculation of the gradient at the cell edge, and by the overall geometry of the cell, the ratio of cell perimeter to cell area. This metric can vary from 0.0 to 3.0 or more depending upon the geometric features of the cell, although most images of cells have values between 0.0 and 1.0. Figure 2 shows three cells with very different extended edge neighborhoods. If the edge is very thick because the image is blurry, as in the third picture of Figure 2, the extended edge neighborhood can be larger than 1.0. If a cell is very thin but the cell edges are very sharp, as in the middle picture of Figure 2, it still has a higher extended edge neighborhood than a larger cell with similar edges, as in the first picture of Figure 2. This metric can be calculated for each cell in the absence of good segmentation, because knowledge of the cell edge is not required for either the edge thickness or cell geometry estimates. The calculation for each cell can be performed efficiently even for very large datasets.

5 Testing 40000 Cells

Figure 3 shows a histogram plot of the EEN metric for 40000 cells. It shows that the A10 cells have lower EEN values on average than the NIH3T3 cells. To determine if the extended edge neighborhood metric is predictive of accuracy

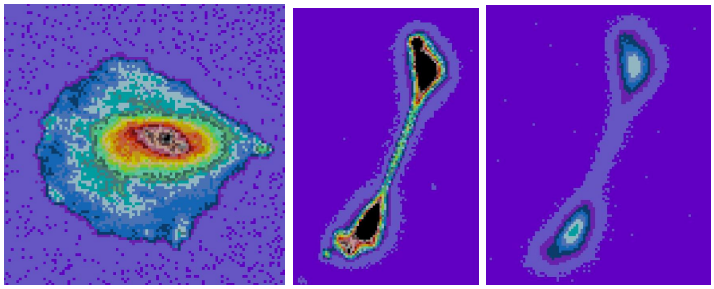

Fig. 2. 3 cells are colored according to pixel intensity, with the full range shown divided into 40 different colors: a large, round cell with low extended edge neighborhood; a small, thin cell with higher extended edge neighborhood, even though the edges are shar; the same cell but a blurrier image, and the extended edge neighborhood is greater than 1.0.

for segmentation algorithms, we studied the accuracy of four segmentation algorithms that use different methods to identify cell edges as a function of extended edge neighborhood for each cell. The algorithms tested were 3-means clustering, 4-means clustering, 5-means clustering, and a Canny edge method. Segmentation masks were generated from the k-means clustering algorithms by assuming that the cluster with the lowest centroid represents the background and the remaining clusters belong to the cells. To determine an accuracy metric for each cell that is segmented by an automated segmentation, we compared the results of the algorithm to that of a reference segmentation data set derived with an a computer assisted manual segmentation and expert visual inspection, using bivariate similarity metrics, previously described in [7] [9]. Definitions of these metrics and a justification for their use are summarized in the next section.

6 Bivariate Similarity Index

Various similarity metrics have been used to evaluate segmentation algorithm performance. The commonly used Jaccard similarity index [13], for example, compares a reference data set, T, with another set of estimates, E, defined by:

$$S = |T \cap E|/|T \cup E|, \tag{2}$$

where $0.0 \leq \mathrm{S} \leq 1.0$. If an estimate matches the truth, $\mathrm{T} \cap \mathrm{E} = \mathrm{T} \cup \mathrm{E}$ and $\mathrm{S} = 1$. If an algorithm fails, then $\mathrm{E} = 0$ and $\mathrm{S} = 0$. However, S cannot discriminate between certain underestimation and overestimation cases. For example, if the true area = 1000, then both the underestimated area of 500, and the overestimated area of 2000 yield the same value for the similarity index $\mathrm{S} = 500/1000 = 1000/2000 = 0.5$. Here we used a set of bivariate similarity indices that can distinguish between underestimation and overestimation.

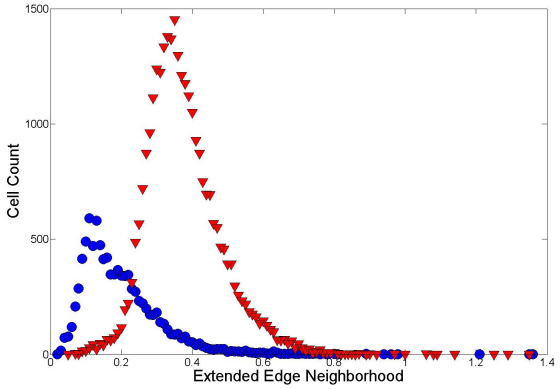

Fig. 3. Numbers of cells as a function of extended edge neighborhood for the A10 cells in red, and the NIH3T3 cells in blue.

We define these indices as follows, to compare the reference pixel set T, with a segmentation mask, pixel set E:

$$TET = |T \cap E|/|T|, 0.0 \leq TET \leq 1.0 \tag{3}$$

$$TEE = |T \cap E|/|E|, 0.0 \leq TEE \leq 1.0 \tag{4}$$

Each similarity metric varies between 0 and 1. If the estimate matches the reference mask, both TET and TEE = 1.0. TET and TEE were constructed to be independent and orthogonal and divides performance into four regions: Dislocation: TET and TEE are small; Overestimation: TET is large, TEE is small; Underestimation: TET is small, TEE is large; and Good: both TET and TEE are large. Figure 5 illustrates the use of the indices to compare a group of approximately 3000 A10 cells, segmentated with a 5-means clustering segmentation in Figure 4. This plot shows the tendency of 5-means clustering segmentations to underestimate cell edge compared to the manually segmented data.

In some situations, segmentation comparisons may be facilitated by combining the bivariate indices into a univariate metric. For these purposes we define a metric called the segmentation distance as the Euclidean distance from the point corresponding to the TET and TEE values to the point corresponding to perfect segmentation ($TET = 1.0, TEE = 1.0$). The univariate metric does not contain information about over- or undersegmentation, but it does provide a general measure of segmentation accuracy and can be used to evaluate correlations with the extended edge neighborhood metric.

7 Results

To begin, we evaluate which segmentation algorithm worked best for each individual cell. Figure 5 shows how many of the 40000 cells scored best for each

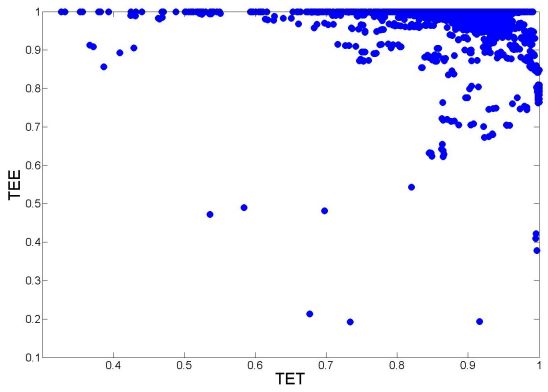

Fig. 4. Plot of TET vs. TEE for 3,000 A10 cells.

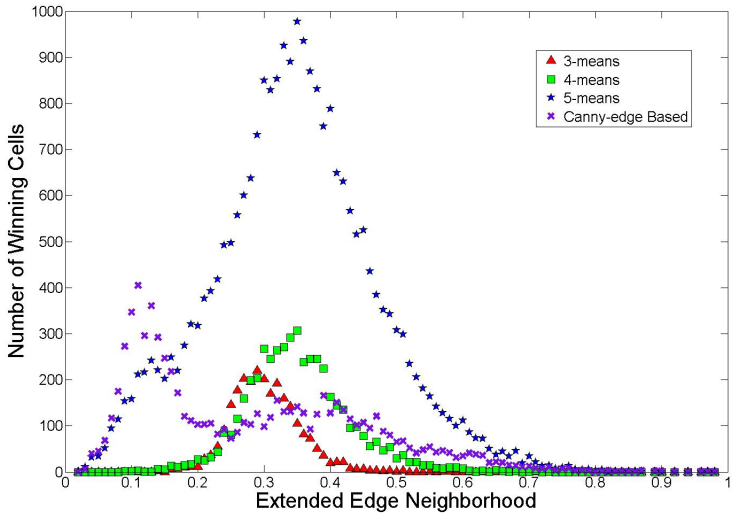

Fig. 5. Number of cells for which a technique worked best: 3-means clustering (red), 4-means clustering (green), 5-means clustering (blue), Canny edge (purple), as a function of extended edge neighborhood.

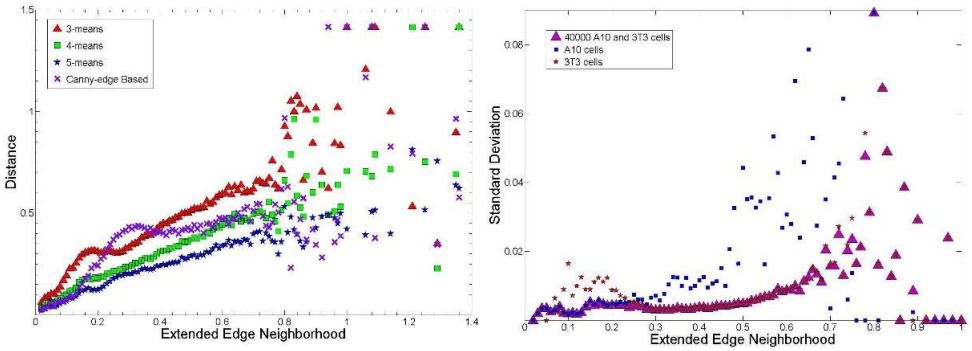

Fig. 6. Averaged segmentation distance for each group of cells with the same extended edge neighborhood for 3-means clustering (red), 4-means clustering (green), 5-means clustering (blue), and Canny edge (purple); Standard deviations of 5-means clustering averaged segmentation distance results as a function of extended edge neighborhood for the A10 cells (blue); for the NIH3T3 cells (red); for all 40000 cells (purple).

method, as a function of our new metric, the extended edge neighborhood. Often several of the methods gave similar results, but this plot counts the number of times each method gave the best results, regardless of whether another method came close. If two methods gave identical results for a given cell, results for that cell were not included in the plot. The 5-means clustering gives the best results for this cell image dataset and the cells most likely to be segmented best using this method tended to have a mean EEN value greater than 0.2. Interestingly, the Canny edge segmentation method works very well in only a small region of the extended edge neighborhood curve, at low extended edge neighborhoods between 0.0 and approximately 0.15, which represents larger cells with sharp edges. In this region the Canny algorithm segmentation results are more similar to manual segmentation than 5-means clustering, 4-means clustering, or 3-means clustering. 3-means clustering and 4-means clustering methods did best for only a small number of cells in the extended edge neighborhood region between 0.2 and 0.4, where the accuracy using any method was not very high.

Figure 6 shows all of the segmentation results for each of four methods over the whole range of our extended edge neighborhood metric. The results are presented in terms of the averaged segmentation distance for groups of cells with the same extended edge neighborhoods. We can draw a number of conclusions from this data. In general, the accuracy of these methods varies monotonically with extended edge neighborhood, and the 5-means clustering results are on average always better than the 4-means clustering, which are always better on average than the 3-means clustering. An occasional cell has better results for 3-means clustering than 4-means clustering or 5-means clustering. However, the results show that in general, the extended edge neighborhood metric predicts the accuracy of k-means algorithms in segmenting these cell images. The standard

deviations of segmentation distances averaged over a group of cells with the same extended edge neighborhoods are low, suggesting a high level of predictability for most cells. As an example, the standard deviation results for the 5-means clustering data are shown in the second plot of Figure 6. The data show that for A10 smooth muscle cells, the standard deviation in segmentation accuracy is low for the EEN range 0.0 to 0.2. The standard deviation for the NIH3T3 cells is low for the range 0.2 to 0.5.

The Canny edge segmentation results are similar to the 5-means clustering method at low extended edge neighborhoods. Above we saw that more cells were segmented accurately in this extended edge neighborhood than with 5-means clustering, but the variability of the Canny edge results produce a similar plot on data averaged over all of the cells, in the region between 0.0 and 0.15 extended edge neighborhood. Overall, we see a fairly linear trend in the best averaged segmentation as a function of extended edge neighborhood.

To investigate further the relationship between the EEN metric, the extended edge neighborhood, and the best averaged segmentation results, we look only at data for which there are a large number of cells. We find the region on the extended edge neighborhood curves for each cell line that includes 90 % of the cell data in each cell line. The first two plots of Figures 7 overlay plots of cell counts for the A10 and NIH3T3 cell lines respectively with the segmentation distances from a 5-means clustering for each cell line. In the third plot of Figure 7, we graph the A10 cells in the extended edge region from 0.02 to 0.2, and the NIH3T3 results in the extended edge region 0.2 to 0.5, where 90% of the cells from each cell line occur. All of the data for cells whose extended edge neighborhoods are less than 0.5 are fitted with a linear model, which is also plotted in Figure 7: predicted distance = 0.051 + EEN $\times$ 0.477, with a correlation coefficient of 0.9815.

8 Conclusions and future work

From this large scale test, we define a method to pre-process images and determine their vulnerability to segmentation error. The accuracy that is possible from any given segmentation technique is directly proportional to the extended edge neighborhood of each individual cell within an image. Rounder, larger cells have a lower extended edge neighborhood than smaller less round cells, and segmentation will more closely align with manual segmentation for these cell images. Our results suggest that of the four segmentation methods tested here, a 5-means clustering segmentation is the most reliable. The Canny edge segmentation method performs best with cells within a very small extended edge neighborhood range that is less than approximately 0.15. We can now use the methods outlined in this paper to look at a wider range of segmentation algorithms for identifying more accurate segmentation techniques. We have written a software segmentation pre-processor that calculates extended edge neighborhood for each cell in an image and then provides the best technique and expected accuracy for the segmentation of each cell based on these four algorithms, which

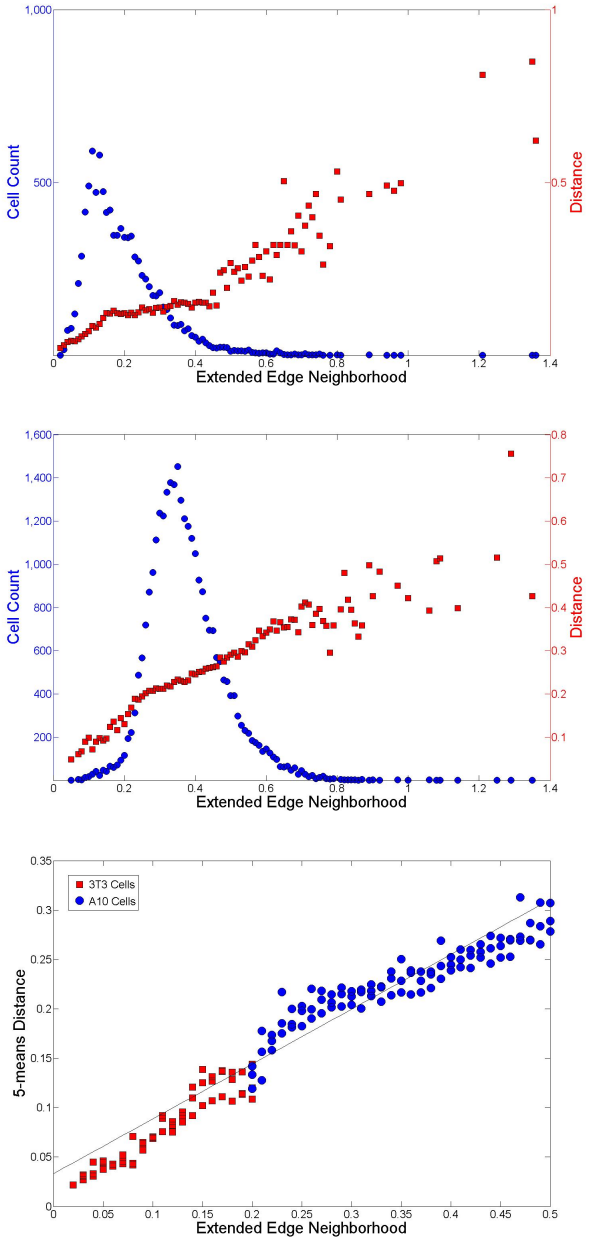

Fig. 7. Averaged segmentation distance for the A10 cells (red), along with a plot of cell numbers as a function of extended edge neighborhood (blue); Same for the NIH3T3 cells; Results from the A10 cells (red) and NIH3T3 cells (blue) from the first two plots. A straight line is fitted to this data.

can evolve as we compare more algorithms. We believe that this processor will be of great use for optimizing the segmentation of cells seeded at low density and stained as described here. The EEN can also be used as a metric to rank cells for most-likely best segmentation. Measurements of cell function can be weighted by this ranking to potentially improve the measurement robustness in a cell-based assay. The EEN metric will have significant value in determining which cells in a data set are most at risk during a segmentation procedure.

9 Acknowledgements

References

1. Plant, A.L., Elliott, J.T., Tona, A., McDaniel, D., Langenbach, K.J.: Tools for Quantitative and Validated Measurements of Cells. High Content Screening: A Powerful Approach to Systems Cell Biology and Drug Discovery. Humana Press, Taylor,L., Giuliano,K., and Haskins,J., ed. (2006)
2. Elliott,J.T., Tona,A. Plant,A.L.: Comparison of reagents for shape analysis of fixed cells by automated fluorescence microscopy. Cytometry. 52A:90-100 (2003)
3. Elliott,J.T., Woodward,J.T., Langenbach,K.J., Tona,A., Jones,P.L., Plant,A.L.: Vascular smooth muscle cell response on thin films of collagen. Matrix Biol. 24(7), 489-502 (Oct 2005)
4. Zhou X., Wong, S.T.C.: High content cellular imaging for drug development. IEEE Signal Processing Magazine. 23(2), 170–174 (March 2006)
5. Coelho, L.P.,Shariff, A., Murphy, R.F.: Nuclear Segmentation in Microscope Cell Images: A Hand-Segmented Dataset and Comparison of Algorithms. ISBI (2009)
6. Cardinale, J., Rauch, A., Barral,Y., Szkely, G., Sbalzarini, I.F.: Bayesian image analysis with on-line confidence estimates and its application to microtubule tracking. IEEE International Symposium of Biomedical Imaging. 1091-1094 (June 2009)
7. Dima,A.,Elliott,J.T.,Filliben,J.,Halter,M.,Peskin,A.,Bernal,J., Stotrup,Marcin,Brady,A.,Plant,A.,Tang,H.: Comparison of segmentation algorithms for individual cells. Cytometry Part A. in process.
8. Langenbach, K.J., Elliott, J.T., Tona, A., and Plant, A.L.: Evaluating the correlation between fibroblast morphology and promoter activity on thin films of extracellular matrix proteins. BMC-Biotechnology 6(1):14 (2006)
9. Chalfoun,J., Dima, A., Peskin, A.P.: A New Protocol for Generating Semi-Automatic Generation of Reference Data for High Resolution Biological Images, in process.
10. Peskin, A.P., Kafadar, K., Dima, A.: A Quality Pre-Processor for Biological Cells. 2009 International Conference Visual Computing (2009)
11. Hastie, T., Tibshirani, R.; Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, And Prediction. New York, Springer (2001)
12. Peskin, A.P., Kafadar, K., Santos, A.M., Haemer, G.G.: Robust Volume Calculations of Tumors of Various Sizes. 2009 International Conference on Image Processing. Computer Vision, and Pattern Recognition (2009)
13. Rand,W.M.: Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association. 66(336), 846-850 (Dec 1971)