

# Comparison of Segmentation Algorithms For Fluorescence Microscopy Images of Cells

Alden A. Dima,<sup>1</sup> John T. Elliott,<sup>2</sup> James J. Filliben,<sup>3</sup> Michael Halter,<sup>2</sup> Adele Peskin,<sup>4</sup> Javier Bernal,<sup>5</sup> Marcin Kocielek,<sup>6</sup> Mary C. Brady,<sup>1</sup> Hai C. Tang,<sup>1</sup> Anne L. Plant<sup>2\*</sup>

<sup>1</sup>Software and Systems Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899

<sup>2</sup>Biochemical Science Division, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899

<sup>3</sup>Statistical Engineering Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899

<sup>4</sup>Applied and Computational Mathematics Division, Information Technology Laboratory, National Institute of Standards and Technology, Boulder, Colorado 80305

<sup>5</sup>Applied and Computational Mathematics Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899

<sup>6</sup>Technical University of Lodz, Medical Electronics Division, Lodz, Poland

Received 19 August 2010; Revision Received 24 February 2011; Accepted 12 April 2011

Additional Supporting Information may be found in the online version of this article.

## • Abstract

The analysis of fluorescence microscopy of cells often requires the determination of cell edges. This is typically done using segmentation techniques that separate the cell objects in an image from the surrounding background. This study compares segmentation results from nine different segmentation techniques applied to two different cell lines and five different sets of imaging conditions. Significant variability in the results of segmentation was observed that was due solely to differences in imaging conditions or applications of different algorithms. We quantified and compared the results with a novel bivariate similarity index metric that evaluates the degree of underestimating or overestimating a cell object. The results show that commonly used threshold-based segmentation techniques are less accurate than k-means clustering with multiple clusters. Segmentation accuracy varies with imaging conditions that determine the sharpness of cell edges and with geometric features of a cell. Based on this observation, we propose a method that quantifies cell edge character to provide an estimate of how accurately an algorithm will perform. The results of this study will assist the development of criteria for evaluating interlaboratory comparability. Published 2011 Wiley-Liss, Inc.<sup>1</sup>

## • Key terms

fluorescence microscopy; k-means cluster; image segmentation; cell edge; bivariate similarity index

**NUMEROUS** areas of biomedical research rely on imaging of cells to provide information about molecular and phenotypic responses of cells to pharmaceuticals, toxins, and other environmental factors (1,2). Cell imaging is widely used in biological experiments because it can provide information on several relevant scales simultaneously. Molecular and supramolecular scales can be probed with the use of antibody and other specific affinity reagents and with fluorescent proteins. Gross phenotypic characteristics of cells that characterize their ultimate functional state can be examined in the same experiment, and often the temporal regime can be probed simultaneously. Together, these applications of cell imaging allow inference about the molecular details and the complex outcomes of the cellular biochemistry.

Because of the enormous number of parameters that may influence a biological outcome, cell imaging experiments are often done in a “high content” mode (3,4), where large numbers of paired and replicate experiments are carried out simultaneously and result in very large (often gigabyte) image datasets. Such a large volume of image data precludes visual inspection of every image, and automated image processing and analysis is the only viable approach to data analysis.

Segmentation of cell objects is a common image analysis operation that provides spatial and other features of identified objects and often precedes other operations to quantify parameters such as intracellular fluorescence. Segmentation can pose significant challenges to automated image processing and analysis. Because morphological features are often important indicators of a complex cellular response, and because

\*Correspondence to: Anne Plant, Biochemical Science Division, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899

Email: anne.plant@nist.gov

Published online 14 June 2011 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cyto.a.21079

Published 2011 Wiley-Liss, Inc.<sup>†</sup>

This article is a US government work and, as such, is in the public domain in the United States of America.

accurately quantifying intracellular fluorescence may be dependent on having an accurate cell mask provided by a segmentation operation, segmentation, or edge detection is often a critical image analysis operation.

Many software packages are available that provide image processing and analysis tools, and many microscope and high content screening instrument manufacturers provide proprietary software. Segmentation algorithms are often used as 'black boxes', and assumptions are made about their accuracy. There have been several studies published on comparing segmentation algorithms for cells (5–7). A study that examined how poor segmentation of cellular images can influence the interpretation of image data was published by Hill et al. (8). The study examined the analysis of SK-BR-3 cells in a high content screening assay to compare experimental treatments that affected cell morphology. Their study determined that 64% of cells were poorly segmented by the algorithm used, resulting in a failure to detect the effect of the treatment, an effect that became apparent when only well-segmented cells were considered for analysis. Thus the accuracy of segmentation can critically influence experimental conclusions.

In this study, we compare nine different segmentation methods over a range of segmentation types: histogram-based methods, *k*-means clustering methods, and nonhistogram based methods. We test three common histogram-based segmentation algorithms that are available through widely used imaging software such as ImageJ (9–11): Otsu (12,13), isodata (14), and maximum entropy (15). We also examine 2-, 3-, 4-, and 5-means clustering (16). For comparison with a nonhistogram-based approach, we employ an implementation of the Canny edge detector (13,17), as well as the use of a single global threshold value chosen by a microscopist for a given cell line and image condition. We apply these methods to experimental data from two types of mammalian cell lines that have quite different morphologies, A10 vascular smooth muscle cells and NIH 3T3 fibroblasts. A10 cells tend to be larger and better spread than NIH 3T3 cells, and have smoother edges. The methods are applied to fluorescence images of fixed cells that have been covalently labeled by treatment with a common fluorophore reagent. The cells were seeded at a low density to ensure high contrast between the cells and the background at the periphery of each cell. We compare datasets containing the same images collected under different acquisition conditions that result in variation in the ratios of background intensity to cell intensity and the sharpness of focus. The outcome of the segmentation was determined by comparing the results with reference data generated by manually outlining cell objects. While the data used in this study do not

visually appear to be very difficult to segment, there is a surprisingly large variation in the results achieved from the different algorithms. In this report, we present quantitative metrics by which to compare algorithm performance, and offer an analysis of what cell image features are likely to be problematic for which algorithms.

## MATERIALS AND METHODS<sup>1</sup>

### Cell Culture

A-10 rat smooth muscle cells and NIH-3T3 mouse fibroblasts (ATCC, Manassas, VA) were maintained in Dulbecco's Modified Eagles Medium (DMEM/10 %FBS, Mediatech, Herndon, VA) supplemented with glutamine, nonessential amino acids and occasionally penicillin/streptomycin (Invitrogen, Carlsbad CA) in 5% CO<sub>2</sub> at 37°C. For the experiment, the cell lines were seeded at 800 and 1,200 cells/cm<sup>2</sup>, respectively, in three-wells of a six-well tissue culture treated polystyrene plate (353046, BD Falcon, Franklin Lakes, NJ) in maintenance media, and placed in the incubator for ~20 h. The media was removed; the cells were rinsed with PBS and fixed for 3 h with 1% (v/v) formaldehyde in PBS at 25°C. The cells were stained with PBS containing 0.02% (v/v) TritonX-100 (Sigma, St. Louis, MO), 0.5 μg/mL TxRed c2 maleimide (Invitrogen) (5 mg/mL in DMSO stock), 1.5 μg/mL DAPI (Sigma) (1 mg/mL in DMSO stock) for 4 h, rinsed with PBS, PBS containing 1% BSA and PBS, sequentially. Fixed and stained cells were covered with PBS, stored at 4 °C, and imaged within two days.

### Automated Fluorescence Microscopy Imaging

Fluorescence images of fixed and stained cells were acquired with an Olympus IX71 inverted microscope (Center Valley, PA) equipped with an automated stage (Ludl, Hawthorne, NY), automated filter wheels (Ludl), a Xe arc fluorescence excitation source, a 10 x ApoPlan 0.4 NA objective (Olympus), and a CoolSNAP HQ charge-coupled device (CCD) camera (Roper Scientific, Tucson, AZ). The filter combinations (Chroma Technologies, Brattleboro, VT) were as follows: the optimal filters for imaging the TxRed stained cells were a 555 nm ± 12.5 nm band pass excitation filter (PN# S555\_25x) and a 630 nm ± 30 nm band pass emission filter

<sup>1</sup>Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

**Table 1.** Acquisition conditions and parameters

IMAGE CONDITION	EXPOSURE TIME (S) A10	EXPOSURE TIME (S) 3T3	FILTER TYPE <sup>a</sup>	SNR <sup>b</sup>	RESOLUTION <sup>c</sup> (lp/mm)
1	0.015	0.01	optimal filter (555 nm excitation, 630 nm emission)	25±8	203
2	0.08	0.05	optimal filter (555 nm excitation, 630 nm emission)	103±31	203
3	0.3	0.15	optimal filter (555 nm excitation, 630 nm emission)	221±72	203
4	1.0	1.0	non-optimal filter (555 nm excitation, 740 LP emission)	20±9	114
5	5.0	5.0	non-optimal filter (555 nm excitation, 740 LP emission)	50±21	114

<sup>a</sup> Filters present in filter wheels. A multipass dichroic beam splitter (BS51019, Chroma Technology) is used for all fluorescence imaging.

<sup>b</sup> The SNR was calculated as described in the Materials and Methods.

<sup>c</sup> Indicates the smallest line features (lp/mm=line pairs per millimeter) that could be resolved in the 1951 USAF target under the optical settings used for cell imaging (see Supporting Information 1). These values correspond to group 7 element 5 (203 lp/mm) and group 6 element 6 (114 lp/mm) line patterns of the resolution target.

(PN#S630\_60m). For the non-optimal filter conditions, the emission filter was replaced with a custom 740 nm long pass filter. For nuclear imaging of the DAPI stain, a 360 nm ± 20 nm band pass excitation filter (PN#D360\_40x) and a 460 nm ± 25 nm band pass emission filter (PN#D460\_50m) were used. For each condition, light passed through a custom coated multipass dichroic beam splitter (PN#51019+400DCLP) matched to these filters. A fluorescent glass artifact (Schott GG475, Edmund Scientific, NJ) was imaged with an FITC filter set to adjust lamp alignment to minimize spatial variations in illumination, and to assess day-to-day stability in lamp intensity (18). No flat field correction was performed on the images. The Xe lamp intensity through the dichroic beam splitter (with no excitation filter) at the exit site of the 10× objective was measured to be approximately 171 mW/cm<sup>2</sup> in a 1 cm<sup>2</sup> area detector (Newport Instruments, Stratford, CT).

The microscope stage, CCD and automated shutters were controlled by modular routines within ISee image acquisition software (ISee Imaging Systems, Raleigh, NC). For each well of the three wells in each plate, a grid of 50 non-overlapping fields were imaged. At each field, the following operations were performed sequentially: (1) the optimum filters were set for TxRed detection, the excitation light shutter was opened, (2) auto-focusing was performed, (3) three consecutive images of the TxRed stained cells were acquired at each of three different exposure times (see Table 1 for exposure times), (4) the filters were changed to nonoptimal conditions and three images were collected at each of two different exposure times, (5) the filters were switched for DAPI detection and a nuclear image was acquired (2 s integration time), (6) the emission shutter was closed and the transmitted light shutter was opened, (7) a phase contrast image was acquired and the transmission shutter was closed, (8) the stage was moved to the next position. All images were acquired using 2 x 2 binning on the CCD sensor. Only the TxRed images were used for segmentation analysis.

A large dataset was collected, consisting of 500 images of 100 different fields, to allow for future studies to investigate the effects of shot noise, of well-to-well variability, and of day-to-day variability. During acquisition, day-to-day variability was tested by imaging the same fields after storing the samples at 4°C overnight. A fiducial mark was made at the corner of

a well by crossing two pieces of transparent tape. The stage position coordinates were zeroed at this location and fields of view were nearly identically reproduced during the image collection routine. We concluded from preliminary experiments that day-to-day acquisition variability, well-to-well variability and variation due to shot noise) were negligible sources of uncertainty compared to uncertainty due to image analysis.

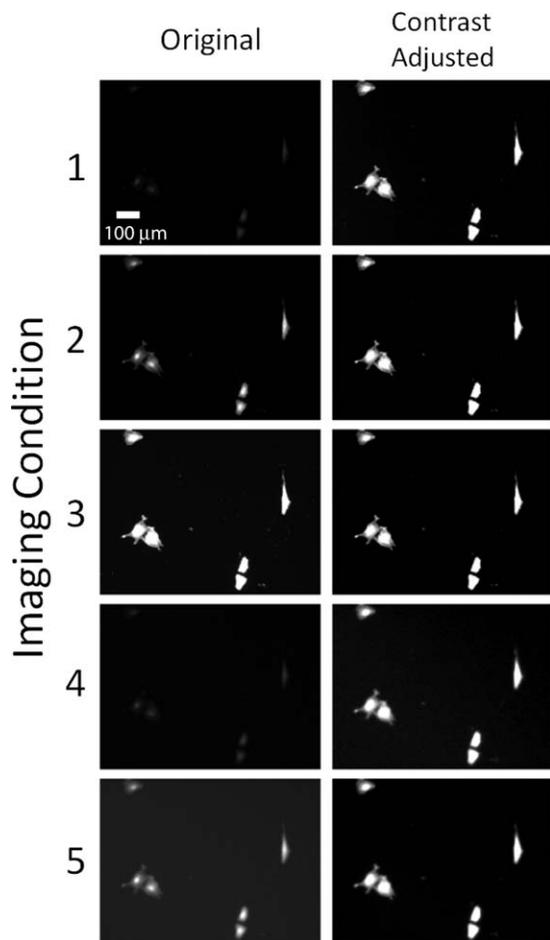
For our current purposes, we selected a representative subset of 16 representative fields, eight from each cell line, based on their representation of a range of cell sizes and cell densities. Each field of cells was represented by images acquired under five different image acquisition conditions, for a total of 80 images. A summary of the imaging conditions used is given in Table 1, and a set of representative images are shown in Figure 1 for the five conditions.

### Parameterization of the Image Collection Settings

We used the following procedures and Image J image analysis software (NIH, <http://rsbweb.nih.gov/ij/>) to parameterize the five imaging conditions. Variations in exposure times were parameterized as average signal-to-noise ratios (SNR). Using object masks from the manual segmentation reference data, the image signal was calculated as the average intensity of cell-associated pixels over all images for that imaging condition, minus the average intensity of the noncell pixels, and the noise was determined as the standard deviation of the intensity in the non-cell pixels. These values are shown in Table 1. Applying the non-optimal filters resulted in a change in the SNR and a degradation of the sharpness of the image. This degradation was quantified as a decrease in spatial resolution as determined using a chromium coated 1951 USAF resolution target (Edmund Optics, Barrington, NJ) with a fluorescent Schott OG650 glass filter (Edmund Optics, Barrington, NJ) above it (Supporting Information Fig.™ 1). The smallest line features that could be resolved are listed in Table 1 as line pairs per millimeter.

### Reference Data

The images acquired under imaging condition 3 had the brightest and most distinct cell edges and therefore were used for generating the manually segmented reference masks. Cell objects in these images were delineated by manually outlining



**Figure 1.** Representative images of A10 cells showing the effect of image acquisition conditions. See Table 1 for details of imaging conditions.

cells or groups of cells using the Paintbrush tool in ImageJ with a brush width of one pixel. Cells in contact with one another were treated as a single object. To facilitate the accurate identification of the cell boundary the image of each cell was enlarged with the Zoom tool. Masks were generated from the outlines of the cell images using the ImageJ particle analyzer. To test the variation in the manual segmentation data generated with this procedure, two different individuals generated manually segmented masks. One set of manually segmented masks was selected as the reference, and the other manual segmentation data and the algorithm results were compared with it.

### Algorithm Implementation

The sixteen representative fields imaged under the five imaging conditions comprised a set of 80 images contained a total of 71 unique cell objects, 31 from the A10 cell line and 40 from the NIH 3T3 cell line. The cell objects in these images were segmented using the 9 different methods listed in Supporting Information Table 1: global threshold (glo), Canny edge detection (cn), 2-means clustering (k2), 3-means cluster-

ing (k3), 4-means clustering (k4), 5-means clustering (k5), maximum entropy (mx), isodata (iso), and Otsu thresholding (ot). The resulting masks were compared to the reference segmentation masks generated by manual segmentation using the bivariate metric described below. Supporting Information Table 1 also reports information about the algorithm variables that can be modified to change the segmentation results. The isodata, Otsu, and maximum entropy methods were implemented as part of the ImageJ Multithresholder plug-in and were used without alteration. The k-means clustering and Canny edge-based segmentation were implemented in FORTRAN and called via an ImageJ plug-in that managed their execution in a separate operating system process.

The *k*-means implementation initialized the clusters by randomly selecting values from the range of pixel values found in an image instead of the more typical strategy of randomly selecting pixel values directly from the image itself. The implementation also produced cluster maps in which the pixels were labeled with cluster IDs sorted by centroid values. This cluster initialization and sorting strategy minimized some of nondeterministic features associated with *k*-means clustering and led to consistent results of repeated trials. The first cluster, which had the lowest centroid value, was assumed to be the background.

The Canny edge-based segmentation followed the approach outlined in Gonzalez and Woods (13,17). First the input image was smoothed with a circular 2D Gaussian filter. The gradient magnitude and angle images were then computed from the smoothed image so that non-maxima suppression could be applied to the gradient magnitude image using the angle image. Finally, the edges were detected and linked using hysteresis thresholding and connectivity analysis.

The results from all of the segmentation methods were post-processed using a morphological threshold to eliminate particles from debris that was determined to be too small to be cellular objects. This morphological threshold value varied by cell line; for the smaller NIH 3T3 cells the threshold was set at 50 pixels and for the larger A10 cells it was set at 200 pixels. These thresholds were set based on experience that objects of this size and smaller are not cell objects. This criterion was used in order to assure robust, unsupervised, automated segmentation. Cells were excluded from the analysis if their corresponding reference data mask touched an image boundary to reduce the ambiguity associated with partial cell objects.

Segmentation methods occasionally returned a number of cell objects in an image that was different from the number of cell objects in the reference data image. Missing cells and cells which fragment during segmentation can lead to a different labeling of the same cell across segmentations and make comparisons difficult and time-consuming. To facilitate the automated comparisons of segmentation results, an overlap-based region-of-interest (ROI) converter was implemented in MATLAB and was invoked prior to further analysis. This was accomplished by first processing the binary output mask from each segmentation method by the ImageJ particle analyzer to output a mask (via the Count Mask feature) in which each object's pixel values identifies the object. The ROI converter

then relabeled the objects to be consistent with the reference segmentation. Afterwards, additional MATLAB routines were invoked to analyze the relabeled masks.

### Analysis pipeline

The evaluation was run in an unattended fashion using ImageJ macros to apply each segmentation method to every image in the dataset. The ImageJ macros themselves were created by a set of code generators written in Perl to facilitate their maintenance and extension.

After an evaluation run, the segmentation masks were compared with the reference data masks in MATLAB. The results of the segmentations were compared with manually outlined cell objects to evaluate the algorithm performance.

### Extended Edge Neighborhood Calculation

We developed a metric to explore the role that the gradient of the pixel intensity at the cell edge can play in determining whether a cell image can be properly segmented. These concepts have been combined into a single quantity that can be calculated quickly for each cell in an image. The extended edge neighborhood (EEN) is a fraction derived from the ratio of pixels at risk of being misclassified to the total area of the cell. The pixels at risk of misclassification are determined from the cell perimeter and a quality index (QI) described below. The cell edge thickness is then determined from the quality index.

For each cell in an image, its quality index is found from the pixel intensities within an isolated region containing the cell. These steps, described in detail in (19), are summarized as:

1. Find the 3-component Gaussian mixture via the EM (Expectation-Maximization) algorithm, whose components correspond to background ( $B$ ), edge ( $E$ ), and cell ( $C$ ) pixels, and denote the means of the components by  $\mu_B$ ,  $\mu_E$ , and  $\mu_C$ .
2. Find the average gradient magnitude at each intensity between  $\mu_B$  and  $\mu_E$  and denote the resulting function by  $G(\text{Intensity})$ . Because not all possible intensity values are present in the portion of the image containing the cell, there will be gaps where no average gradient magnitude can be calculated. These missing average gradient values are approximated by interpolation.
3. Find the intensity,  $A$ , at which the smoothed average gradient magnitude is maximized.
4. Since intensity  $A$  has the maximum average gradient magnitude, all pixels adjacent to  $A$  will have a lower average gradient magnitude. The pixel intensities also decrease across the cell edge from the cell interior to the background. We can calculate the expected intensity  $B$  of a pixel neighboring  $A$  closer to the cell edge, and pixel  $B$  will have a lower gradient magnitude as well as a lower intensity value. Calculate the expected intensity  $B$  of a neighboring pixel to a pixel with intensity  $A$  as

$$B = A - G(A) \times \Delta x, \text{ where } \Delta x = 1 \text{ pixel unit.}$$

5. Calculate the expected intensity  $C$  of a pixel neighboring a pixel with intensity  $B$ ; i.e.,

$$C = B - G(B) \times \Delta x = A - G(A) \times \Delta x - G(A - G(A) \times \Delta x) \times \Delta x.$$

6. Calculate the quality index as  $QI = (A - C)/(A - \mu_B)$ .

The quality index ranges from 0.0 to 2.0, with a perfectly sharp edge at a value of 2.0, which would correspond to an edge thickness of 1.0 pixel unit. Figure 2 has an example of two A10 cells acquired under our five imaging conditions colored to show the various resulting edge thicknesses. Therefore, to find the edge thickness in general, we define edge thickness ( $T$ ) and then extended edge neighborhood (EEN), as:

$$T = 2.0/QI$$

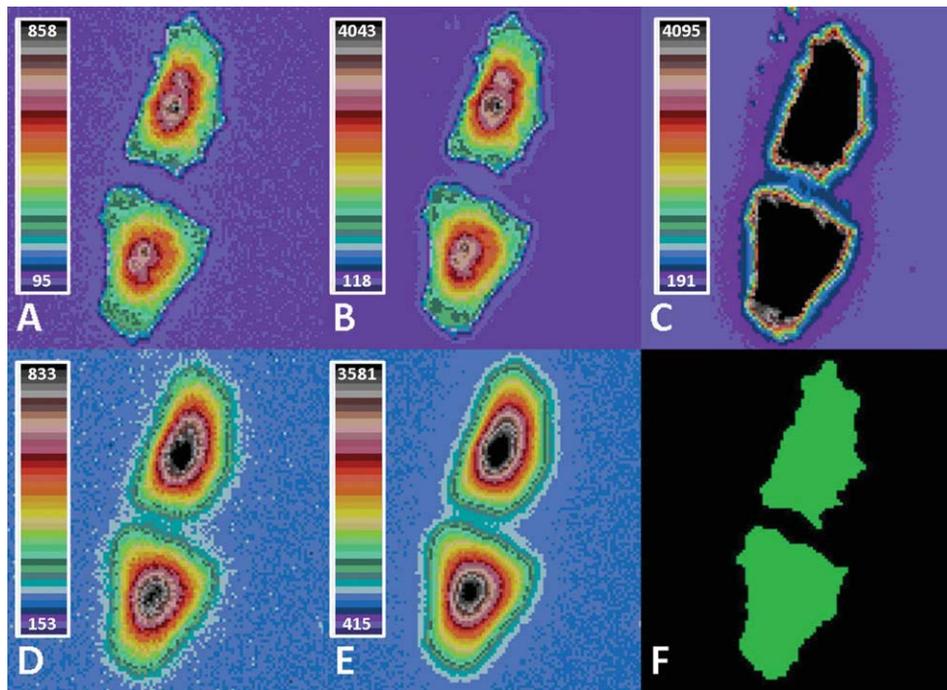
$$\text{EEN} = \text{perimeter} \times T / \text{area}$$

The extended edge neighborhood is defined by the intensity contrast of the image, through the calculation of the gradient at each cell edge, and by the overall geometry of the cell via the ratio of cell perimeter to cell spread area. For a more detailed description of this index, see (20).

## RESULTS

### Bivariate Similarity Index

A range of approaches for evaluating biomedical segmentation algorithms appear in the literature. Visual inspection of the segmentation algorithm results is perhaps the simplest and most intuitive; it is also widely used (21–27). Another approach is the use of a metric which quantitatively assesses the performance against reference segmentations. A variety of metrics have been used. Gulda et al. (28) and McCullough et al. (29) use root-mean-square (RMS) deviations to evaluate their segmentation algorithms while Li et al. (30) and Russell et al. (31) use false positive and negative rates. Several approaches based on set theory also appear. Zijdenbos et al. (32) use the Jaccard similarity index based on set intersection and union operations to examine segmentations of brain lesions, while Ko et al. (33) compare their nuclear segmentations to manual segmentations by defining over- and under-extraction ratios based on set difference and intersection operations which they then use to calculate the segmentation accuracy. Korzynska et al. (34) take a similar approach when defining a cell area error metric for their live cell segmentation method as does Srinivasa et al. (35) when they use area similarity to evaluate their active mask algorithm for fluorescent microscopy. Recently, Polak et al. (36) have proposed the object-level consistency error (OCE) which is a sophisticated set theory-based metric. Statistical approaches to segmentation algorithm evaluation are similar to set theory-based approaches and include work done by Coelho et al. (6) to evaluate nuclear segmentations using statistical indices, Cardinale et al.'s (37) use of Bayesian image analysis to validate their



**Figure 2.** A–E: Two A10 cells acquired under five imaging conditions (1–5, respectively). The manual reference segmentation for the cells is shown in (F). For each pair of cells (A–E) the pixel intensity range of the image was divided in 40 equal smaller ranges. The 40 colors from low (purples) to high (black) are shown in the color table. The manual reference segmentation for the cells is shown in (F).

framework for subcellular feature segmentation and tracking, and the proposed use of the Normalized Probabilistic Rand (NPR) index by Unnikrishnan et al. (38).

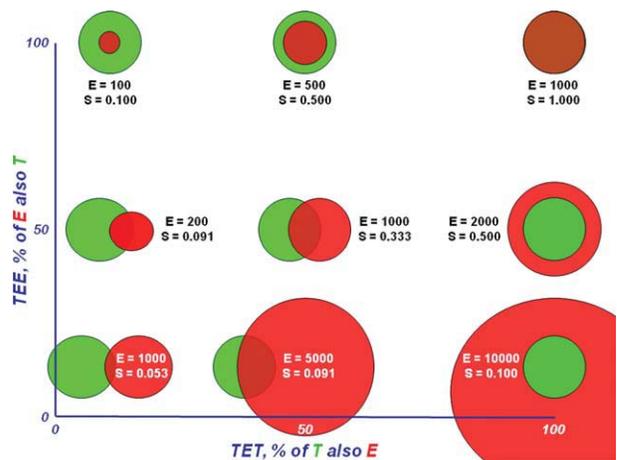
The results of the more mathematically sophisticated approaches can be difficult to interpret in practice. While the performance of segmentation algorithms can be evaluated by simple criteria such as measured cell areas with respect to reference values, such criteria do not adequately allow assessment of how algorithms fail. Another approach is the use of common similarity indices, but because of ambiguities that are described below, for this analysis we will present a bivariate similarity index.

The commonly used Jaccard similarity index (32,39) which compares a reference data set with another set of estimates  $E$  is defined as:

$$S = \frac{|T \cap E|}{|T \cup E|} \text{ where } 0 \leq S \leq 1.$$

For our purposes,  $T$  represents the pixels in a reference mask (“truth”) and  $E$  represents the pixels in a mask produced by a segmentation technique (“estimate”). The numerator represents the count of the pixels in common while the denominator represents the total number of pixels in both masks. If an estimate matches the reference then  $T \cap E = T \cup E$  and  $S = 1$ . If an algorithm fails to segment a cell object then  $E = 0$  and  $S = 0$ . However,  $S$  cannot discriminate between certain underestimation and overestimation cases. Figure 3 schematically demonstrates this dilemma. For exam-

ple, if the reference and the estimate for an object are both circular and concentric, then the underestimation case where the reference area  $T = 1000$  and the estimated area  $E = 500$  yields a similarity  $S = 500/1,000 = 0.5$ . A corresponding overestimation case with  $\Delta T = 1,000$  and  $E = 2,000$  also yields a similarity  $S = 500/1,000 = 0.5$ . Despite their different situations, both of these cases yield the same similarity index value. Because the degree of under-versus-over estimation is an im-



**Figure 3.** A schematic framework for bivariate similarity index.  $T$  (green circles) represents the pixels associated with the object ( $T = 1,000$ );  $E$  represents the estimated pixels from a segmentation procedure. For  $E = 10,000$ , only a portion of the red circle is represented.

portant characteristic of a segmentation algorithm, this similarity index ambiguity is a disadvantage, suggesting the need for developing an alternative similarity index that would provide such directional over-versus-under estimation information.

We propose the use of a bivariate pair of similarity indices defined as follows:

$$TET = \frac{|T \cap E|}{|T|} \text{ where } 0 \leq TET \leq 1$$

and

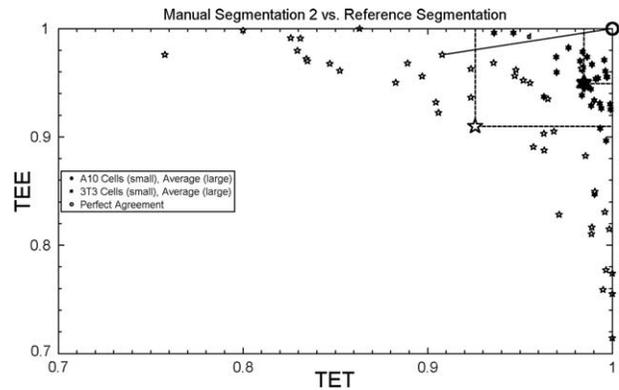
$$TEE = \frac{|T \cap E|}{|E|} \text{ where } 0 \leq TEE \leq 1.$$

As with the similarity index described above,  $T$  represents the pixels in a reference mask (“truth”),  $E$  represents the pixels in a mask produced by a segmentation technique (“estimate”) and the numerators represent the count of the pixels in common. Here however, the denominators represent the total number of pixels in the reference mask and or a given segmentation mask, respectively. If the estimate matches the truth then  $T = E$  and  $TET = 1$  and  $TEE = 1$ . If an algorithm fails to segment an object then

$E = \phi$   $TET = \phi$  and  $TEE$  is undefined. For the purposes of this paper, we defined  $TEE = 1$  when  $E = \phi$  as this corresponds intuitively to the case of a threshold being set too high for thresholding methods.

$TET$  and  $TEE$  are symmetrically defined, bounded, and sufficiently independent and orthogonal for our purposes. They serve as a valid mathematical basis for evaluating segmentation algorithms, and we believe that they provide more information and easier interpretation than previously used metrics.  $TET$  and  $TEE$  are very similar to precision and recall from precision-recall framework used to evaluate boundary-based segmentations of natural scenes (40–42). Both are bivariate means of comparing segmentation quality. Precision and recall also have forms which are similar to our  $TEE$  and  $TET$ , respectively. There are some key differences between the two approaches. The precision-recall framework uses an affinity function to calculate mutual information between the segmentations containing multiple classes. Precision and recall are defined in terms of shared classification agreements between all possible pixel pairs in both segmentations (40). Our  $TET/TEE$  bivariate metric assumes only two classes (foreground and background) and directly compares corresponding pixels from each segmentation.

$TET$  and  $TEE$  are also similar to the under- and over-extraction ratios proposed by Ko et al. (33) for use in calculating their segmentation accuracy metric but are made simpler by removing the set difference operations. This bivariate metric divides performance into four regions: (1) where  $TET$  and  $TEE$  are both small, i.e., the object and estimate are dislocated from one another as in Fig. 3, lower left, (2) where  $TET$  is large and  $TEE$  small, i.e., the object is overestimated as in Fig. 3, lower right, (3) where  $TET$  is small and  $TEE$  large, i.e., the object is underestimated as in Fig. 3 upper left, and (4) where



**Figure 4.** TET vs. TEE bivariate similarity index plot for the two manually segmented sets of each of the 71 cell objects collected over the five image conditions used in this study. Results for the A10 cells are shown as black six-pointed stars, the results for NIH 3T3 cells as white five-pointed stars. The segmentation distance is shown for one pair of segmentations.

both  $TET$  and  $TEE$  are large, i.e. the object is well segmented as in Fig. 3, upper right.

In some situations, segmentation comparisons may be facilitated by using a univariate metric instead of a bivariate metric. For this these purposes, we have defined the segmentation distance as the Euclidean distance from the point corresponding to perfect segmentation ( $TET = 1$ ,  $TEE = 1$ ):

$$d_{\text{seg}} = \sqrt{(1 - TET)^2 + (1 - TEE)^2}$$

The segmentation distance retains the disadvantages of the similarity index in that it does not uniquely define the type of segmentation failure for any comparison. For example,  $d_{\text{seg}} = 1$  when  $TET = 1$  and  $TEE = 0$  as well as when  $TET = 0$  and  $TEE = 1$ .

### Comparison of Manual Segmentations

Figure 4 shows the use of the bivariate similarity index to compare the results of the two sets of manual segmentations. As described in the Methods section, of the two individual sets of manually generated segmentation masks for the 71 cell images, one was arbitrarily chosen as the reference for all other segmentations. If manual segmentation #2 for a cell was a perfect segmentation relative to the reference mask for that cell, it would be displayed at the center of the open circle at the point (1.0, 1.0) on the plot. The comparison of the manual segmentation of each A10 cell with its corresponding reference manual segmentation is indicated by its Euclidean distance from the circle and is shown as black six-pointed stars with the average indicated by the larger black six-pointed star. The corresponding data for NIH 3T3 cells are shown as white five-pointed stars. It is apparent in this figure that the two manual segmentations are more consistent for the A10 cells than the NIH 3T3 cells. Comparison of the results of the two manual segmentations indicates that there appears to be no systematic

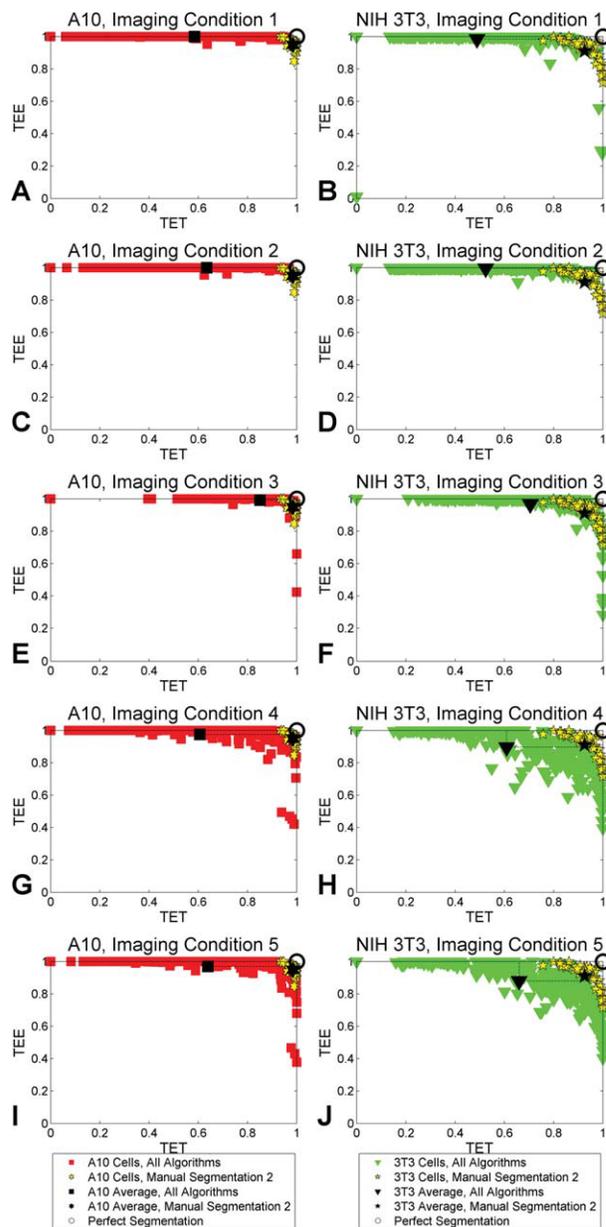
bias between two different humans' interpretations of cell edges, since both under- and over-segmentation occur.

### Comparison of Automated Segmentation Methods

Each cell object mask that resulted from the automated segmentation methods was compared to the manual reference segmentation. Segmentation operations were performed on each cell image collected under the different acquisition conditions. Thus, for each algorithm, we compared masks from the 71 cell objects and 5 different acquisition conditions for a total of 355 comparisons.

In Figure 5 we show all results for all 355 comparisons for each of the nine segmentation methods as a function of imaging condition (see Table 1 for descriptions of imaging conditions). Each segmentation mask was compared with the reference data mask and plotted as a function of TET and TEE with the data for A10 cells in red and the data for NIH 3T3 cells in green. A perfect segmentation, relative to the reference mask, is displayed by an open circle at the point (1.0, 1.0) on the plot. In general, all segmentation methods gave higher accuracy, relative to the reference data, for the larger, well spread A10 cells than for the smaller NIH 3T3 cells. For both cell lines, the average TET and TEE values are indicated by the red A10 marker and the green NIH 3T3 marker. We show average values of the bivariate statistic because the average is a measure of the central tendency, and is a rational way to compare algorithms to one another on a cell-by-cell basis. The TET and TEE values for the 2nd nonreference manual segmentations are also indicated in Figure 5 by the yellow circles, and the average values by the black A10 markers and blue NIH 3T3 markers. Some segmentation results lie within the region of the plot occupied by 2nd manual segmentation results; such segmentation results are considered to be indistinguishable from manually segmented results. It is interesting to note that unlike the comparison of the manual segmentations, the automated methods demonstrate considerable bias in the inaccuracy of the segmentation. For the larger (A10 cells), especially, inaccuracies that result in low values for TET tend to have large values for TEE, indicating that the estimated cell area is more often underestimated than overestimated. This trend is less pronounced in the case of the 3T3 cells, which have more optical contrast at their edges. These kinds of situations are easy to detect with a suitable bivariate metric but are not discernible with a univariate metric.

The data in Figure 5 demonstrate the effect of image acquisition condition on the performance of these algorithms. The imaging condition that provided the average segmentation that was closest to the reference was condition 3, which provided the highest SNR (as shown in Table 1). Imaging condition 3 resulted in many cell objects that were intensity saturated; this apparently did not adversely affect the histogram-based analyses. Imaging condition 3 also resulted in higher contrast at the cell edge. We will comment on this further in our discussion of the sharpness of cell edges below. As seen in Table 1, conditions 1 and 2 involved shorter exposure times, resulting in smaller SNR; conditions 4 and 5 were additionally



**Figure 5.** Segmentation results by imaging condition for all segmentation methods (with no distinction made between segmentation methods), plotted for each imaging condition. TET and TEE bivariate similarity index plots for all 71 cell objects examined under the five imaging conditions and nine segmentation methods. Data within triangles indicate failure to identify one or more cell objects. See Table 1 for the descriptions of the imaging conditions and Supporting Information Figure 2 for an enlarged version of this figure.

degraded by the use of suboptimal filters, resulting in poorer spatial resolution.

The error in cell segmentation can be large, as seen where points have low TET or TEE values. Determination of reported cell areas led to a mean relative bias of  $-32.5\%$  across all examined automated methods and imaging conditions. Figure 5A–5D shows the effect that imaging conditions can have on segmentation accuracy and why. Many segmented cells imaged

under conditions 1 and 2 have high TEE and low TET values, indicating that the estimate of number of cell pixels is much smaller than the true number of pixels. This result suggests that these segmentation methods do not do a good job in distinguishing cell edges from background, particularly when SNR is low. Increasing SNR (as for imaging condition 3), improves segmentation accuracy, largely by reducing underestimation. Average segmentation accuracy moves closer to perfect segmentation for both cell lines for imaging condition 3. Reducing both spatial resolution and SNR under imaging conditions 4 and 5 resulted in average segmentation accuracies that were similar to those seen for imaging conditions 1 and 2. However, the reduction in spatial resolution under imaging conditions 4 and 5 resulted in algorithm performance that was less biased toward under-estimation. While many cells are still underestimated (e.g. they have a high TEE / low TET values), compared to image acquisition conditions 1 and 2, inaccuracy was more often due to overestimation as well, as indicated by more cells having high TET / low TEE values. Especially for 3T3 cells, many segmentation operations resulted in intermediate TEE/TET values, indicating that cells were inaccurately segmented due to the simultaneous inclusion of background pixels and the failure to include cell pixels. In contrast, for both cell lines, manual segmentation produced much less systematic bias, and much less sensitivity to imaging conditions. It is noted that the two manual segmentations of A10 cells are much more similar than the two manual segmentations of 3T3s.

To better evaluate what appears to be the optimal imaging condition, the plots in Figure 6 show the evaluation of the masks obtained as a function of segmentation technique for image condition 3. Comparing the average results from the nine different segmentation techniques indicates that 4 and 5-means clustering and maximum entropy were the best performing algorithms with respect to concurrence with manual segmentation. Maximum entropy, however, failed to find some cells. User selection of a single global threshold value for a cell line under a given imaging condition also performed very well. The poorest performing algorithms of those examined were isodata, Otsu, and 2-means. TEE/TET analysis shows that failure of these algorithms is largely due to underestimation of cell pixels. The location of the marker for the average segmentation accuracy for isodata indicates the simultaneous contribution from background pixels misclassified as cell pixels as well as cell pixels misclassified as background pixels. Examination of selected images indicated that underestimation of segmented areas was usually the result of exclusion of a large number of cell pixels resulting in a mask with a heavily eroded appearance and segmentation of only the brightest cell body features, resulting in either a single small mask or several disconnected small masks within a cell object. Even under these imaging conditions, some of the algorithms, namely maximum entropy, Canny edge detection, and isodata, failed to find one or more cells, as indicated by points lying within the triangle on some plots.

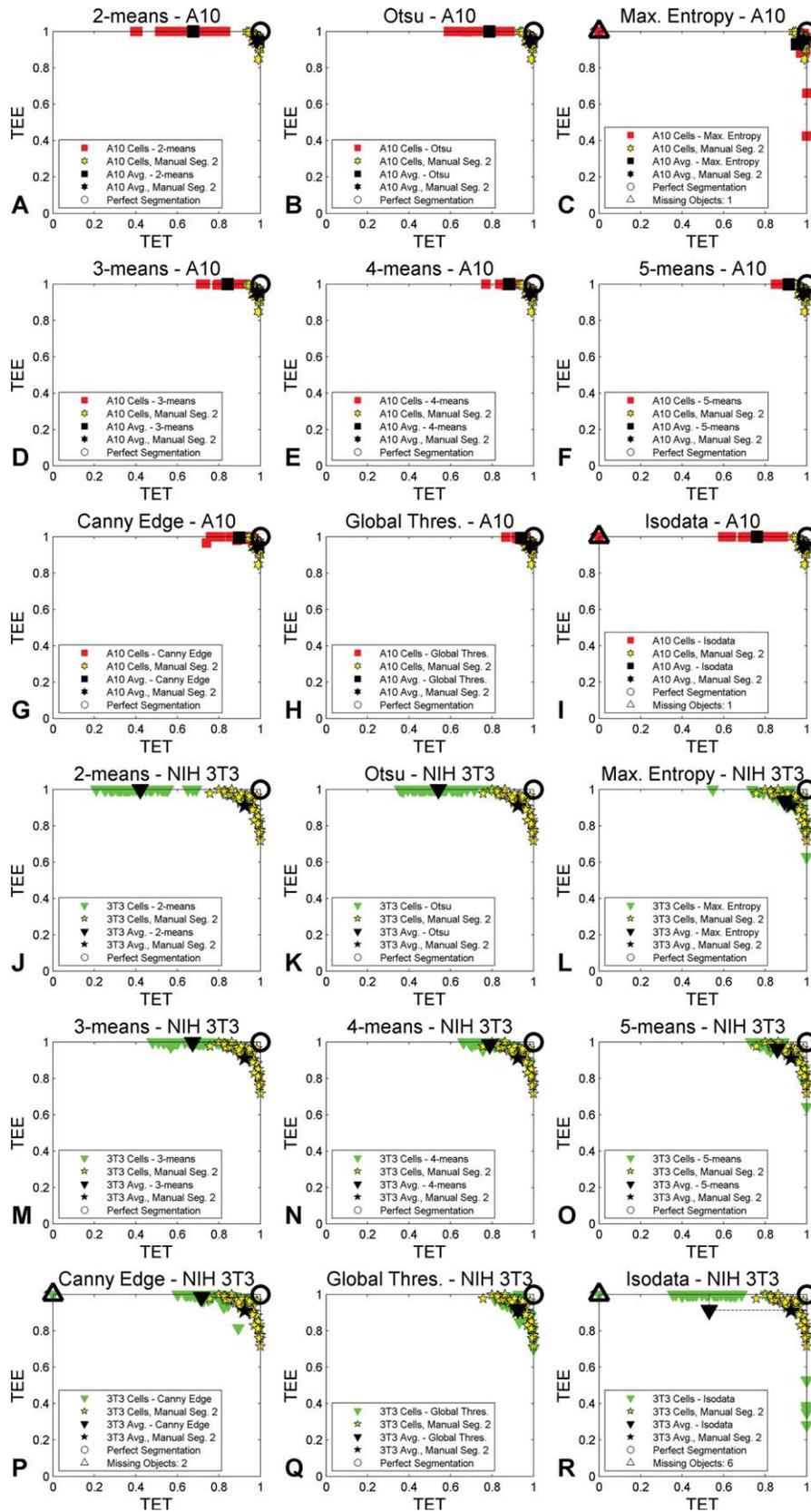
A summary of the relationship between algorithm, imaging condition, and accuracy of segmentation is shown gra-

phically in Figure 7 and numerically in Table 2. The segmentation distance is taken as the Euclidean distance from perfect segmentation ( $TET = 1$ ,  $TEE = 1$ ), where perfect segmentation would have a Euclidean distance of zero. It is clear from the values in Table 2 that no automated method outperforms manual segmentation, even though the two manual segmentations are not identical. Imaging condition 3 resulted in the best concurrence with the manual segmentation reference for all methods. Five-means clustering resulted in the lowest or next to lowest Euclidean distances over all imaging conditions. Maximum entropy, while a poor performer under most conditions, performed as well as 5-means clustering under imaging condition 3, where the SNR was high. Segmentation using isodata gave poor results even under the best imaging condition, condition 3. Many of the algorithms failed to find some cells (as indicated by a Euclidean distance of 1.0); this occurred for even relatively accurate methods such as 4-means clustering, although only under the poorest imaging conditions tested.

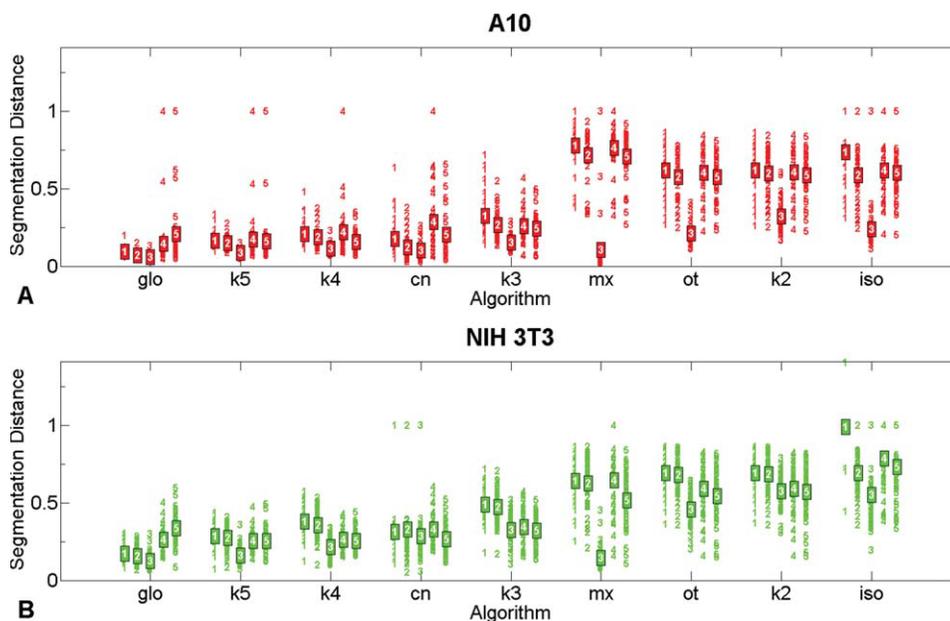
### Effect of Edge Quality on Segmentation

We observed in Figure 5 that imaging condition 3, which resulted in the largest SNR, produced the best segmentation results. Because of the relatively long exposure times, imaging condition 3 produced high object intensities and relatively sharp object edges. We now examine how the characteristics of the cell object edge can potentially influence the success of segmentation. In particular, for each cell object, a value for edge quality was computed based on the steepness of the gradient at the cell edge (19). Cells with edge pixels that are very distinct from background values have a high edge quality, while cells that have shallow gradients of intensities at their edges have a lower edge quality. We combined a measure for edge quality, outlined in a previous article (19), with information about the geometry of each cell, to give an overall measure of the fraction of cell pixels for a given cell which are on or near the edge of the cell and thus at risk to be segmented improperly. We call this measure the extended edge neighborhood, and it is the ratio of the pixels in the neighborhood of the cell edge to the area of the cell. Images of large round cells, like many of the A10 cells we examined, have a high fraction of pixels inside the cell, away from the cell edge, while small, thinner cell objects, like most of the images of the 3T3 cells, have a larger fraction of pixels at risk, near the edge of the cell. Small thin cells with indistinct edges have an even larger fraction of pixels at the edge than small thin cells with sharp edges.

In Figures 8 and 9 we examine the relationship of the extended edge neighborhood (EEN) metric to the results of segmentations of the A10 and NIH 3T3 cells. Segmentation distances are shown for cell images collected under different imaging conditions and analyzed with each of the algorithms. The plots in Figures 8 and 9, in general, demonstrate two concepts. The first is that for all of the segmentations produced with the algorithms evaluated, there is a clear trend between imaging condition and segmentation distance: for images acquired using the optimal filter, the distances are largest for imaging condition 1, intermediate for imaging condition 2,



**Figure 6.** Bivariate similarity plots of segmentation results by segmentation method for both cell lines under imaging condition 3. Data within triangles indicate failure to identify one or more cell objects. See Supporting Information Figure 3 for an enlarged version of this figure.



**Figure 7.** The average segmentation distance for each evaluated algorithm under the five imaging conditions and two different cell lines. The numbers for the data indicate the imaging condition as shown in Table 1. The designation of algorithms is listed in Supporting Information Table 1a. “Glo” refers to a user-selected global threshold. The algorithms have been arranged in order of increasing average segmentation distance for the combined results of all of the imaging conditions. The points marked with a colored box containing a white number represent the mean values for each imaging condition.

and smallest for imaging condition 3. A similar trend holds for the images acquired using the non-optimal filter (imaging conditions 4 and 5). Thus, imaging conditions that tend to result in poorer SNRs tend to lead to cell objects with lower edge quality, and in general, poorer segmentation accuracy. Edge quality is a major factor of our extended edge neighborhood concept. The other factor of the EEN metric is cell geometry, but clearly the metric is relevant to the different shapes of both the cell lines studied here.

In general, we observe that the segmentation distance decreases with decreasing EEN metric. For the algorithms with the poorest accuracy, namely maximum entropy, isodata, and

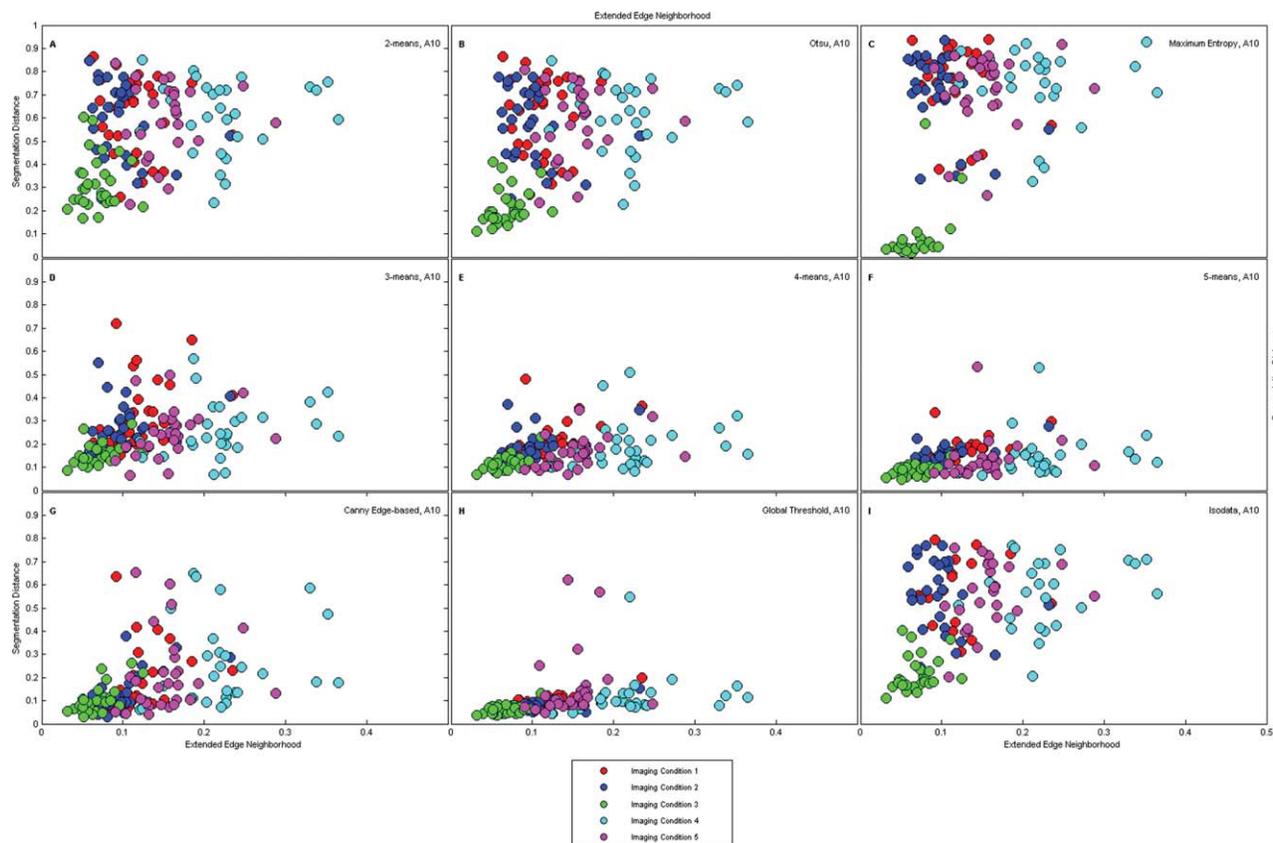
Otsu, EEN is a less clear predictor, suggesting perhaps that the weakness in these methods is less a result of SNR. These techniques typically underestimate the area of a cell which suggests that they select threshold values that are too high. The edges of the resulting segmentation masks are further away from the actual cell edges than those produced by the more accurate methods. As a result, it appears that the EEN, a measure of the cell edge, is a less effective predictor for these methods.

The extended edge neighborhood therefore seems to be a metric that may provide an indication of how well a cell can be segmented. Indeed preliminary results with a larger sampling of the same data, looking at some of the better methods

**Table 2.** Euclidean distances from ideal segmentation for different segmentation methods and imaging conditions

METHOD	CONDITION 1	CONDITION 2	CONDITION 3	CONDITION 4	CONDITION 5	ALL CONDITIONS
k2	0.664	0.651	0.470	0.599	0.581	0.593
Ot	0.663	0.638	0.355	0.598	0.560	0.563
mx	0.702	0.666	0.130	0.697	0.601	0.559
k3	0.422	0.387	0.255	0.311	0.290	0.333
k4	0.310	0.286	0.174	0.247	<b>0.216</b>	0.247
k5	0.237	0.223	0.130	0.221	0.217	0.206
cn	0.256	0.242	0.209	0.312	0.242	0.252
glo	<b>0.141</b>	<b>0.122</b>	<b>0.099</b>	<b>0.215</b>	0.283	<b>0.172</b>
iso	0.885	0.651	0.422	0.718	0.680	0.671
2nd Man. Seg.	—	—	0.107	—	—	—

Euclidean segmentation distances as determined from the bivariate similarity index plot are reported as a measure of segmentation accuracy. The smallest values (bold) represent the highest accuracy with respect to the reference data. See text or Supporting Information Table 1a for algorithm abbreviations. The last row gives the corresponding result from the second non-reference manual segmentation, which was performed on image condition 3 only.



**Figure 8.** Segmentation distance versus extended edge neighborhood for the A10 cells segmented by each algorithm evaluated.

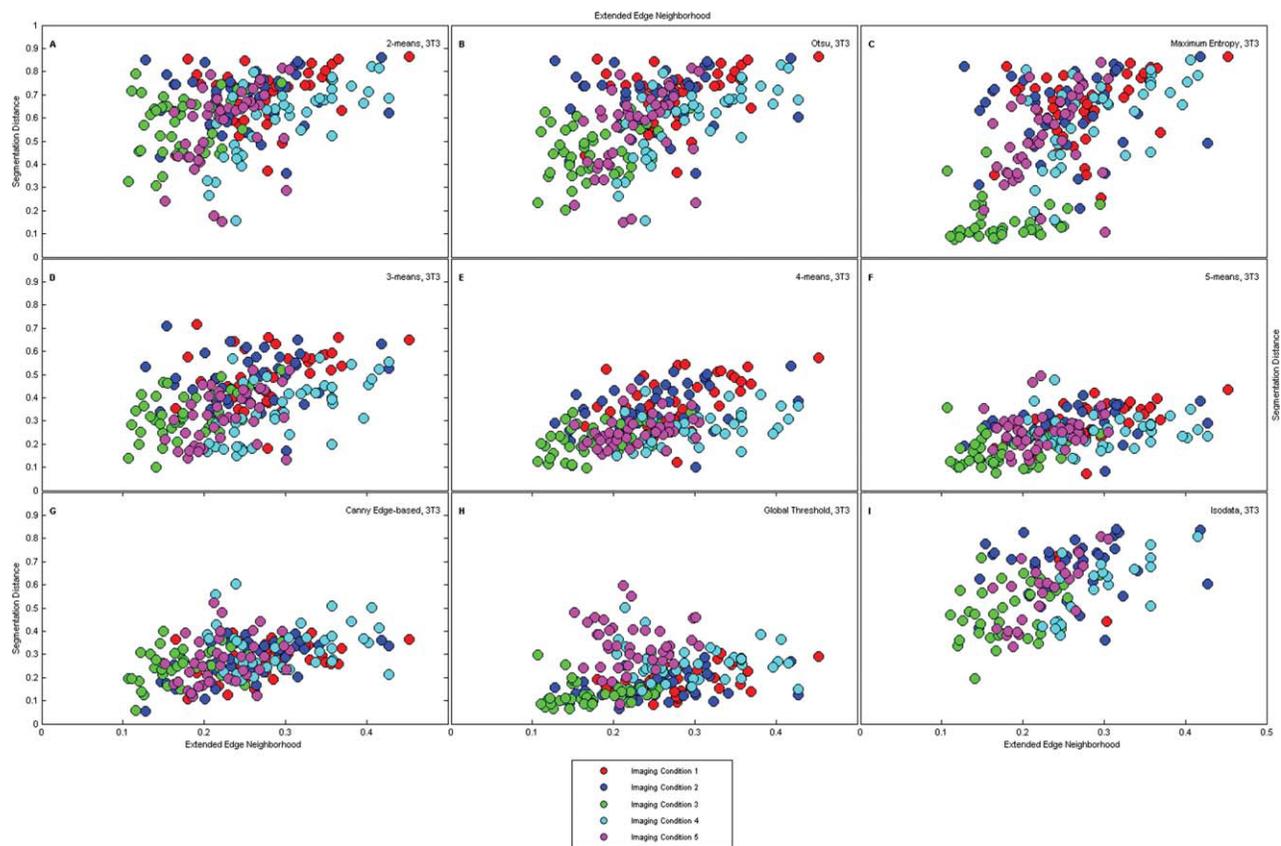
(5-means, 4-means, 3-means, and Canny Edge-based segmentations), show a significant correlation between segmentation distance and extended edge neighborhood. The EEN may also be used to determine whether a cell should be segmented as poorly segmented cells can adversely affect analysis results while retaining only well segmented cells appears to retain biological heterogeneity in certain HCS-unfriendly cell lines (8). Hill et al. (8) trained classifiers to accomplish this task for SK-BR-3 cells but it is possible that setting an EEN threshold could perform this task without having a training set.

## DISCUSSION

The collection of large cell image datasets demands the use of automated image analysis. Because human supervision during analysis is often precluded by the amount of data to be analyzed, the quality of the analysis is dependent on how well the applied algorithms perform. How well any algorithm performs is a function of cell morphology and the imaging parameters for each cell. While reports of new and improved algorithms that solve unique discrimination challenges continuously appear, it is rarer that algorithm performance is evaluated by quantitative means. Some examples of quantitative algorithm comparison include use of the Normalized Probabilistic Rand (NPR) index (38) and methods used in research on face recognition. There are few examples of systematic comparisons of image analysis algorithms for cell

image data; these include a comparison of cell shape analysis (5), a recent report from the Murphy lab on comparing segmentation algorithms (6), a study on predicting confidence limits in segmentation of cells (37) and analysis of segmentation of nuclei (7).

In this study, we examined two different cell lines and five imaging acquisition conditions to compare algorithm performance and assess the affect of image quality on segmentation accuracy. We present a bivariate similarity index method for quantitatively evaluating segmentation method performance. The bivariate analysis method can discriminate between overestimation, underestimation and dislocation of the estimated cell mask relative to the manually determined cell mask. The TET versus TEE plots can indicate both random bias, such as in the case where both overestimation and underestimation occur equally (see Fig. 4 for an example), and systematic bias, such as in the case of uniform overestimation (see Fig. 6B for example). Furthermore, this metric can be reduced to a single value (i.e., a Euclidian distance) for ease of reporting comparisons. Previous use of a similar approach, the precision-recall framework (40–42) has been reported for evaluating segmentation of complex natural scenes and for determining the effect of parameters on algorithm performance. The bivariate statistic presented here will be similarly be useful for tracking reduction in bias in algorithm performance as parameters are adjusted but also has the advantage of being



**Figure 9.** Segmentation distance versus extended edge neighborhood for the NIH 3T3 cells segmented by each algorithm evaluated.

intuitive and efficient to compute for fluorescent microscopy images.

We employed manual segmentation of cells as reference against which all other analyses were compared. We consider the differences between two manually segmented sets of data to bracket the acceptable quality of automated segmentation. The uncertainty associated with manual segmentation is well established (6). While manual segmentation is imperfect, the comparison of the results of automated segmentation relative to two manual segmentation operations allows a basis for evaluating the acceptability of the result to the practitioner. A semiautomated segmentation method is under development that may significantly reduce this uncertainty (43).

Except for the Canny edge detection routine, all segmentations used in this study were histogram based. It is important to note that the histogram based methods do not take into account any spatial information and rely only on pixel intensities to separate the pixels into two classes: foreground pixels belonging to the cell and background pixels belonging to non-cell areas. All segmentation methods applied here except for the Canny edge-based method ultimately result in separating pixels into these two classes separated based on a threshold value. The specifics of the operations that provide these results are responsible for the different outcomes in segmentation.

There are many variations of data preprocessing and algorithm parameters that can be applied and compared. For example, we have applied a log transform to the pixel intensities to these images and segmented them with the eight automated segmentation methods, and these results, which are quite different from analysis prior to log transform, are shown in Supplemental Figure 4. To assist the community in testing and quantitative evaluation of segmentation algorithms, we are making this complete dataset (~2500 tif image files) available at [http://xpdb.nist.gov/image/cell\\_image.html](http://xpdb.nist.gov/image/cell_image.html). The database consists of replicates of 50 fields of cells labeled with Texas Red and DAPI, collected under 5 imaging conditions. Users can download phase contrast and fluorescence cell images, images of masks of manual segmentations, and images of spatial resolution targets. Specific datafiles can be selected by semantic queries using extensive experimental metadata.

We report large differences in the quality of performance of these algorithms and substantial differences in response of algorithms to different imaging acquisition conditions. In general, 3-, 4-, and 5-means clustering appeared to be the best performers over the two cell lines and the five acquisition conditions examined. In general, the use of k-means with greater than 2 clusters is similar to a multiple threshold technique. These techniques allow the pixels to be classified into at least 3 clusters (i.e., background, cell edge, and cell body). This sug-

gests that when the cell edge and cell body pixels are placed in a single classifier, they do not form a cluster of pixels with unique character to separate it from the background by the two cluster algorithms. The multiple cluster techniques effectively reduce the weighting of the cell body pixels in the edge threshold selection process, which results in a more accurate segmentation of the cell. Canny edge detection, while significantly more computationally expensive, did not appear to provide an advantage in overall performance. The brightest pixels in the cell body do not significantly influence the cell edge determination. We find that the accuracy of these nine segmentation methods vary with a factor that we have called the extended edge neighborhood, which describes a combination of the gradient at the edge of a cell and the geometric features of the cell. These observations provide an intriguing possibility that an image quality index could predict the likely uncertainty associated with segmentation of that dataset with a particular algorithm.

In this study we found that our image acquisition condition 3, which was associated with the highest SNRs, often resulted in the best segmentation. A potential disadvantage is that the long exposure time resulted in intensity saturation in many cells, which could result in the loss of useful intensity data. Other image acquisition parameters had a greater or lesser effect on algorithm performance depending on the algorithm. User selection of a single global threshold value was not sensitive to lower SNRs but was very sensitive to poor focus. Poor focus degrades the quality of the segmentation of most of the algorithms examined and seemed to negatively affect the segmentation of fibroblasts even more than smooth muscle cells. This is probably due to the different shapes of the cells and the cell-line-specific differences in intensity gradients at cell edges.

In this study, we do not know the true location of the cell edge and we use manually segmented cells to represent the ideal segmentation. An alternative method for obtaining reference segmentation data is to simulate cellular images (44). However, there are challenges associated with simulating cell images because we do not have detailed models for complex image features such as the intensity distribution at the cell edge and the morphology of the cell boundary. In the future, we are interested in using the images studied here to refine models for simulating fluorescence images of cells. High-quality simulations of fluorescent cell images have several advantages for the development and testing of segmentation algorithms: (1) a variety of cell and image characteristics could be systematically examined, (2) the precise location of the cell edge would be known, and (3) the laborious task of outlining cells would be eliminated.

While the evaluation presented here considers whole cell staining, the methods for evaluating segmentation performance can be applied to a variety of the microscopy imaging problems including nuclear staining or tissue sections. In cases where a set of reference segmentation data is available, the TET vs. TEE plots for can be used to evaluate the performance of automated segmentation algorithms. The plots can also be extended to evaluate 3D segmentation, where the TET vs. TEE

axes would represent discrepancies between the reference segmentation and the test segmentation in volume information. The overall goal of this work is to develop a framework for the evaluation of algorithm performance and the uncertainties associated with the automated analysis of cell microscopy images. The results of this study suggest that it would be prudent to run preliminary measurements and manual segmentation results in order to compare the accuracy of possible image segmentation approaches.

## CONCLUSIONS

While the image data chosen for this work were fairly easy to segment, significant differences in algorithm performance were observed. The results from applying the bivariate similarity analysis suggest that, compared with *k*-means clustering and Canny edge detection, the more commonly used segmentation methods (Otsu, Maximum Entropy, isodata) did a poor job as judged by concurrence with human identification of pixels associated with cells. These methods appeared to select a threshold value that was too high and as a result had a tendency to produce masks that were smaller than those of the corresponding reference masks. We also found that higher SNRs in the image resulted in more accurate segmentation by all algorithms as a result of increasing intensity gradients at cell edges, a component of a metric we refer to as extended edge neighborhood.

We present here a bivariate similarity analysis method that provides an indication of how algorithms fail, and we also present an edge quality metric that may prove useful in predicting image segmentation success. The tools developed for this analysis should have general applicability for future studies.

These results provide some guidance for achieving concurrence among sets of image data collected in different laboratories or at different times. Because algorithm results are sensitive to SNRs and spatial resolution, the conditions under which images are collected, as well as the algorithms applied, will influence the analysis.

## ACKNOWLEDGMENTS

The authors thank Richard Rivello and Joe Chalfoun for their help developing and maintaining the software infrastructure needed to perform the evaluations, and students Asheel Kakkad and Sarah Jaffé for creating the manual segmentation masks. They thank Antonio Cardone and Joe Chalfoun for their assistance as the manuscript neared completion. They also thank Charles Hagwood for contributing to the discussions concerning the analysis of the results.

## LITERATURE CITED

1. Yarrow JC, Feng Y, Perlman ZE, Kirchhausen T, Mitchison TJ. Phenotypic screening of small molecule libraries by high throughput cell imaging. *Comb Chem High Throughput Screen* 2003;6:279–286.
2. Giuliano KA, Wing SC, Curran DP, Day BW, Kassick AJ, Lazo JS, Nelson SG, Shin Y, Taylor DL. Systems cell biology knowledge created from high content screening. *Assay Drug Dev Technol* 2005;3:501–514.
3. Giuliano KA, Haskins JR, Taylor DL. Advances in high content screening for drug discovery. *Assay Drug Dev Technol* 2003;1:565–577.
4. Abraham VC, Taylor DL, Haskins JR. High content screening applied to large-scale cell biology. *Trends Biotechnol* 2004;22:15–22.

5. Pincus Z, Theriot JA. Comparison of quantitative methods for cell-shape analysis. *J Microsc* 2007;227:140–156.
6. Coelho LP, Shariff A, Murphy RF. Nuclear segmentation in microscope cell images: A hand-segmented dataset and comparison of algorithms. *Proc IEEE Int Symp Biomed Imaging* 2009;5193098:518–521.
7. Price JH, Hunter EA, Gough DA. Accuracy of least squares designed spatial fir filters for segmentation of images of fluorescence stained cell nuclei. *Cytometry* 1996;25:303–316.
8. Hill AA, LaPan P, Li Y, Haney S. Impact of image segmentation on high-content screening data quality for SK-BR-3 cells. *BMC Bioinformatics* 2007;8:340.
9. Rasband W. ImageJ [Internet]. U. S. National Institutes of Health, Bethesda, Maryland, USA: 1997. Available at: <http://rsb.info.nih.gov/ij/>
10. Abramoff MD, Magelhaes PJ, Ram SJ. Image processing with ImageJ. *Biophotonics Int* 2004;11:36–42.
11. Sheffield JB. ImageJ, a useful tool for biological image processing and analysis. *Microsc Microanal* 2007;13:200–201.
12. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 1979;9:62–66.
13. Gonzalez RC, Woods RE. *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall; 2008.
14. Ridler TW, Calvard S. Picture thresholding using an iterative selection method. *IEEE Trans Syst Man Cybern* 1978;8:630–632.
15. Kapur JN, Sahoo PK, Wong AKC. A new method for gray-level picture thresholding using the entropy of the histogram. *Comput Vis Graph Image Process* 1985;29:273–285.
16. Jain AK, Dubes RC. *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall; 1988.
17. Canny J. A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell* 1986;PAMI-8:679–698.
18. Plant AL, Elliott JT, Tona A, McDaniel D, Langenbach KJ. Tools for quantitative and validated measurements of cells. In: Taylor DL, Haskins JR, Giuliano KA, editors. *High Content Screening*. New Jersey: Humana Press; 2006. pp95–107.
19. Peskin AP, Kafadar K, Dima AA. A quality pre-processor for biological cell images. In: Bebis G, Boyle R, Parvin B, Koracin D, Kuno Y, Wang J, Pajarola R, Lindstrom P, Hinkenjann A, Encarnacao ML, Silva C, Coming D, editors. *Advances in Visual Computing*. Berlin/Heidelberg: Springer; 2009. pp1051–1062.
20. Peskin AP, Dima AA, Chalfoun J, Elliott JT. Predicting segmentation accuracy for biological cell images. In: Bebis G, Boyle R, Parvin B, Koracin D, Chung R, Hammoud R, Hussain M, Tan KH, Crawfis R, Thalmann D, Kao D, Avila L, editors. *Advances in Visual Computing*. Berlin/Heidelberg: Springer; 2010. pp 549–560.
21. Nilsson B, Heyden A. Segmentation of complex cell clusters in microscopic images: Application to bone marrow samples. *Cytometry Part A* 2005;66:24–31.
22. Baggett D, Nakaya M, McAuliffe M, Yamaguchi T, Lockett S. Whole cell segmentation in solid tissue sections. *Cytometry Part A* 2005;67A:137–143.
23. Kim Y, Brox T, Feiden W, Weickert J. Fully automated segmentation and morphometrical analysis of muscle fiber images. *Cytometry Part A* 2007;71A:8–15.
24. Lin G, Chawla M, Olson K, Barnes C, Guzowski J, Bjornsson C, Shain W, Roysam B. A multi-model approach to simultaneous segmentation and classification of heterogeneous populations of cell nuclei in 3D confocal microscope images. *Cytometry Part A* 2007;71A:724–736.
25. Fenistein D, Lenseigne B, Christophe T, Brodin P, Genovesio A. A fast, fully automated cell segmentation algorithm for high-throughput and high-content screening. *Cytometry Part A* 2008;73A:958–964.
26. Yu W, Lee H, Hariharan S, Bu W, Ahmed S. Quantitative neurite outgrowth measurement based on image segmentation with topological dependence. *Cytometry Part A* 2009;75A:289–297.
27. Woodcroft B, Hammond L, Stow J, Hamilton N. Automated organelle-based colocalization in whole-cell imaging. *Cytometry Part A* 2009;75A:941–950.
28. Gulda P, Nandy K, Collins J, Meaburn K, Misteli T, Lockett S. A high-throughput system for segmenting nuclei using multiscale techniques. *Cytometry Part A* 2008;73A:451–466.
29. McCullough D, Gulda P, Harris B, Collins J, Meaburn K, Nakaya M, Yamaguchi T, Misteli T, Lockett S. Segmentation of whole cells and cell nuclei from 3-D optical microscope images using dynamic programming. *IEEE Trans Med Imaging* 2008;27:723–734.
30. Li F, Zhou X, Zhao H, Wong S. Cell segmentation using front vector flow guided active contours. *Med Image Comput Assist Interv* 2009;12(Part 2):609–616.
31. Russell R, Adams N, Stephens D, Batty E, Jensen K, Freemont P. Segmentation of fluorescence microscopy images for quantitative analysis of cell nuclear architecture. *Biophys J* 2009;96:3379–3389.
32. Zijdenbos A, Dawant B, Margolin R, Palmer A. Morphometric analysis of white matter lesions in MR images: Method and validation. *IEEE Trans Med Imaging* 1994;13:716–724.
33. Ko B, Seo M, Nam J. Microscopic cell nuclei segmentation based on adaptive attention window. *J Digit Imaging* 2009;22:259–274.
34. Korzynska A, Strojny W, Hoppe A, Wertheim D, Hoser P. Segmentation of microscope images of living cells. *Pattern Anal Appl* 2007;10:301–319.
35. Srinivasa G, Fickus M, Guo Y, Linstedt A, Kovacevic J. Active mask segmentation of fluorescence microscope images. *IEEE Trans Image Process* 2009;18:1817–1829.
36. Polak M, Zhang H, Pi M. An evaluation metric for image segmentation of multiple objects. *Image Vis Comput* 2009;27:1223–1227.
37. Cardinale J, Rauch A, Barral Y, Szekely G, Sbalzarini I. Bayesian image analysis with on-line confidence estimates and its application to microtubule tracking. *Proc IEEE Int Symp Biomed Imaging* 2009;1091–1094.
38. Unnikrishnan R, Pantofaru C, Hebert M. Toward objective evaluation of image segmentation algorithms. *IEEE Trans Pattern Anal Mach Intell* 2007;29:929–944.
39. Dunham MH. *Data Mining: Introductory and Advanced Topics*. Upper Saddle River, NJ: Prentice Hall; 2003.
40. Martin DR. An empirical approach to grouping and segmentation [Internet]. 2003; Available at: <http://www.eecs.berkeley.edu/Pubs/TechRpts/2003/CSD-03-1268.pdf>
41. Martin D, Fowlkes C, Malik J. Learning to detect natural image boundaries using local brightness, color, and texture cues. *Pattern Anal Machine Intell IEEE Trans* 2004;26:530–549.
42. Fowlkes C, Martin D, Malik J. Learning affinity functions for image segmentation: Combining patch-based and gradient-based approaches. In: *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. Madison, WI: IEEE; 2003.
43. Chalfoun J, Dima AA, Peskin AP, Elliott JT, Filliben JJ. A human inspired local ratio-based algorithm for edge detection in fluorescent cell images. In: Bebis G, Boyle R, Parvin B, Koracin D, Chung R, Hammoud R, Hussain M, Tan KH, Crawfis R, Thalmann D, Kao D, Avila L, editors. *Advances in Visual Computing*. Berlin/Heidelberg: Springer; 2010. pp23–34.
44. Lehmussola A, Selinummi J, Ruusuvoori P, Niemisto A, Yli-Harja O. Simulating fluorescent microscope images of cell populations. In: *Conf Proc IEEE Eng Med Biol Soc*. 2005; pp3153–3156.