Evaluation of Segmentation Algorithms on Cell Populations Using CDF Curves

Charles Hagwood*, Javier Bernal, Michael Halter, and John Elliott

***Abstract*—Cell segmentation is a critical step in the analysis pipeline for most imaging cytometry experiments and evaluating the performance of segmentation algorithms is important for aiding the selection of segmentation algorithms. Four popular algorithms are evaluated based on their cell segmentation performance. Because segmentation involves the classification of pixels belonging to regions within the cell or belonging to background, these algorithms are evaluated based on their total misclassification error. Misclassification error is particularly relevant in the analysis of quantitative descriptors of cell morphology involving pixel counts, such as projected area, aspect ratio and diameter. Since the cumulative distribution function captures completely the stochastic properties of a population of misclassification errors it is used to compare segmentation performance.**

***Index Terms*—Cell morphology, cumulative distribution function (CDF), flow cytometry, image cytometry, misclassification errors, segmentation.**

I. Introduction

IMAGE cytometry is a valuable tool for understanding cellular responses to pharmaceuticals and environmental toxins, as well as discovering correlations between signaling pathways and cell phenotype [1], [2]. A typical analysis often involves imaging cells stained with multiple probe molecules, including a whole cell body stain such as phalloidin or Texas Red c2-maleimide. The whole cell body stain can be used to provide constrast in the image pixels and reveal morphological features such as projected spread area and shape [3], [4]. The cell pixels can also be used to define a mask to integrate a reporter signal (i.e., fluorescent antibody or GFP protein) and may be used to indicate the level of intracellular protein concentration.

In an imaging cytometry experiment, the first step following image acquisition is typically segmentation, where cell and noncell (i.e., foreground and background) pixels are identified and grouped into cell objects. Choosing a reliable segmentation scheme for cell images from a large number of readily available segmentation schemes can be challenging and is crucial for extracting accurate information about cellular features. Through applications, such as NIH ImageJ, numerous implementations of segmentation algorithms are widely available. However, methods for choosing an algorithm that go beyond visual inspection can provide useful tools in the algorithm selection process and in justifying the application of an algorithm for a particular data set.

Any evaluation procedure requires choosing appropriate assessment criteria. A segmentation algorithm for cell images can be considered a pixel classifer and a misclassification results from the algorithm incorrectly labeling a pixel as belonging to background or cell given the observed intensity matrix. Segmentation routines that result in the lowest misclassification errors are often more ideal algorithms [5]. Also, pixels with the same classification are grouped together to form a cell object. The ability to group cell pixels and the misclassification errors can be related processes, but it is also possible for two algorithms to have similar misclassification errors, but result in significantly different cell objects due to differences in how the pixels are grouped. This can lead to a fragmented object. To score the grouping nature of a segmentation algorithm, a fragmentation error can be assessed. Both of these metrics should be taken into account for minimal evalution and comparison of segmentation algorithms.

A unique feature of an imaging cytometry experiment is that information about a population of cells on a cell-by-cell basis can be acquired. If a sufficiently large number of individual cells are sampled in the data set, a robust estimate of the distribution of misclassification errors for the cells in the population can be generated. This is different from other applications of segmentation (e.g., face/object recognition) where the sample size is often not sufficiently large to make inferences about the distribution of misclassification errors. If ground truth segmentation is known for the cells in the cytometry images (i.e., expert manual segmentation) then the percentage $(p, 0 \leq p \leq 100)$ of misclassified pixels per cell in an image due to a segmentation algorithm can be found. Each algorithm will have a unique distribution of misclassification percentages (p's) for any particular cell population. Each misclassification distribution is uniquely characterized by its cumulative distribution function (CDF) $F(p)$. An estimate of $F(p)$ is used to compare and evaluate segmentation algorithms. If there are enough data, then their sample empirical CDFs are good estimators of the population CDFs, since they are maximum likelihood estimators, as well as consistent and unbiased. Evaluation based on misclassification of pixels is particularly important, because controlling and assessing misclassification of pixels are important criteria for the statistical

Manuscript received August 15, 2011; accepted September 12, 2011. Date of publication September 29, 2011; date of current version February 03, 2012. *Asterisk indicates corresponding author.*

*C. Hagwood is in the Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (e-mail: hagwood@nist.gov).

J. Bernal is in the Mathematical and Computational Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (e-mail: bernal@nist.gov).

M. Halter and J. Elliott are in the Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (e-mail: michael.halter@nist.gov; john.elliott@nist.gov).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2011.2169806

analysis of quantitative metrics involving pixel counts which are highly sensitive to the accuracy of the segmentation scheme, such as, projected cell area, aspect ratio, and diameter.

$F(p)$ has several desirable properties as an evaluation criterion. 1) $F(p)$ is an increasing function in p. 2) If $F_A(p)$ and $F_B(p)$ are two CDFs for algorithms A and B, then A is better than B at percentage p if $F_A(p) > F_B(p)$. This is because, if $F_A(p) > F_B(p)$, then the probability that the misclassification percent of A is less than p is greater than the probability the misclassification percent of B is less than p. Therefore, using this criteria, algorithm A can be said to be the best algorithm if its CDF $F_A(p)$ satisfies $F_A(p) > F_B(p)$ for all $0 \leq p \leq 100$ and for all other algorithms B. Typically, it is difficult for uniformity to hold over all $0 \leq p \leq 100$. Alternatively, if a majority of the cells have misclassification percentages less than p_0, then the criterion, algorithm A is the best if $F_A(p) > F_B(p)$ for all $0 \leq p \leq p_0$ and for all other algorithms B may be used. The cutoff p_0 can be found by finding that quantile of their probability density functions (PDFs) where the majority of the mass lies below. When uniformity does not hold in either case, one may rely on other population statistics to help make a decision, such the mean, shape and spread of probability density functions. It is important to note, that in addition to assessment of misclassification errors, the fraction of ground truth cell objects that are fragmented incorrectly also provides information about the performance of an algorithm. Using this criteria, the best algorithms would have a low fraction of fragmented cells. These two quantitative metrics can aid in the decision of algorithm selection which ultimately depends on the experimenter.

We are also interested in the details of how algorithms rank in terms of getting cell shape characteristics correct, such as area, roundness and roughness. This is done by comparing these shape characteristics to those of the manually segmented ground truth cells.

Several previous comparison studies have been performed for cell segmentation, for example in [6]–[11]. Coelho *et al.* [6] compared six thresholding methods including Otsu, a watershed method, an active masks method and a region merging algorithm. Most of these methods are based on summary indexes. The novelty of our approach is that we go beyond summary indexes. Our conclusions are based on an entire population of misclassification percentages. Each algorithm is identified with its population of misclassification errors and comparisons are based on CDFs and familiar population statistics.

In this study, we compared four popular algorithms used for cell segmentation, Otsu thresholding, k-means clustering (with five means), the watershed and the Canny edge detector. Five means have been shown to be near optimal [12]. Two cell types, A10 rat smooth muscle cells and NIH-3T3 cells at three exposures, short, medium, and long are compared. Fifty images of fixed and fluorescently labeled cells that are seeded at low density on a substrate are compared. An image from each cell type at three exposures is shown in Fig. 1.

II. Materials and Methods

This section describes the cells, the image preparation process, the imaging, and the ground truth.

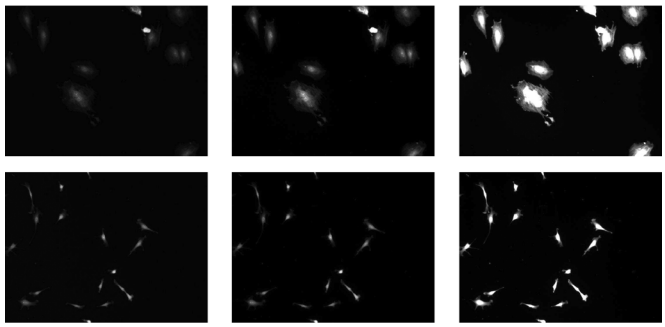

Fig. 1. Cell images at short, medium, and long exposures. Upper panel: A10 rat vascular smooth muscle. Lower panel: NIH 3T3 mouse fibroblasts.

A. Cell Culture

A10 rat smooth muscle cells and NIH-3T3 cells were maintained under DMEM/10%FBS supplemented with glutamine, nonessential amino acids, and occasionally penicillin/streptomycin in 5% CO2 at 37 °C. For the experiment, the cell lines were seeded at (800 and 1200) cells/cm^2 in three-wells, respectively, of a six-well tissue culture treated polystyrene plate (Falcon 4095) in maintenance media and placed in the incubator for approximately 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) TX-100, 0.5 ug/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 containing 0.01% (w/v) sodium azide ([24]). Fixed and stained cells were covered with PBS and imaged within two days.

B. 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 Xenon arc lamp fluorescence excitation source, a 10× ApoPlan 0.4 NA objective, and a CoolSNAP HQ CCD camera (Roper Scientific, Tucson, AZ). The filter combinations (Chroma Technologies, Brattleboro, VT) for imaging the TxRed stained cells were a 555 nm notch excitation (PN# S555_25x) and a 630 nm notch emission filter (PN# S630_60m) with a custom coated multipass dichroic beam splitter (PN# 51019+400DCLP) matched to these filters. The illumination variations in the field of view were minimized using a fluorescent Schott GG475 glass artifact (Edmund Scientific, Barrington, NJ) on a six-well culture plate with an opening in a well bottom, a FITC filter set and lamp alignment and focus adjustments, Plant *et al.* [13]. This focus-based field flattening minimizes the variability in the image segmentation results across the image. The microscope stage, CCD, and automated shutters were controlled by modular routines within ISee image acquisition software (ISee Imaging Systems, Rayleigh, NC). For each well of the six-well plate, a grid of 50 nonoverlapping fields were imaged. All images were acquired using 2×2 binning on the CCD sensor. Three consecutive images of the stained cells at different exposure times were acquired (see Table I for exposure times). Images with varied exposure times

TABLE I
ACQUISITION CONDITIONS AND PARAMETERS. [a]THE SNR WAS CALCULATED AS DESCRIBED IN THE METHODS AND MATERIALS

Image Condition	Illumination Level	Exposure **Time (s)** A10	Exposure **Time (s)** 3T3	Filter Type[a]	SNR ratio[a]	Resolution[a] (lp/mm)
1	Low	0.015 (Short)	0.01 (Short)	optimal filter	25±8	203
				(555 nm		
				excitation, 630		
				nm emission)		
2	Medium	0.08 (Medium)	0.05 (Medium)	optimal filter	103±31	203
				(555 nm		
				excitation, 630		
				nm emission)		
3	High	0.3 (Long)	0.15 (Long)	optimal filter	221±72	203
				(555 nm		
				excitation, 630		
				nm emission)		

were converted to a signal-to-noise ratio using ground truth masks. The signal was calculated as average intensity of cell pixels minus the average intensity of the noncell pixels, and the noise was determined as the standard deviation of the intensity in the noncell pixels. The calculations for all images in the dataset were generated with the use of macros in ImageJ.

C. Evaluation Criteria and Notation

The misclassification data was acquired as follows. Given segmentation and ground truth masks of an image, the goal is to compare cells in an algorithm mask to cells in the ground truth mask. An algorithm may fragment some of the ground truth cells into pieces. Since truth is assumed to be the manual segmentation, these fragments are assumed to be parts of some ground truth cell. Since our cells are seeded at low density, fragments either share common pixels with only one ground truth cell or no ground truth cell. If a fragment has at least one pixel in common with a ground truth cell, then it is assumed to be a part of that cell. So, there is a one to one correspondence between cells in the segmentation mask and cells in the ground truth mask. There are fragments having no pixels in common with a cell in the ground truth mask. This can happen in the identification of cell fragments that are extensions (e.g., pseudopods) of a larger cell. Fragments that cannot be grouped are considered to be totally false and are given a 100% false positive error rate. We keep track of the amount of fragmentation that occurs with an algorithm and report this number as the fraction of cells fragmented in the population.

Before misclassification calculations are performed, we used a 50 pixel size discrimination filter to remove very small groups of cell objects detected in the image. This size filter removed debris that was detected as a cell object and very small pseudopods that likely extend from the cell, but the connection was not readily visable in the image. For the A10 and NIH3T3 cells the 50 pixel filter resulted in a less than 2% reduction of the average cell area and did not significantly change the results of the analysis.

The misclassification rate is calculated as follows. Given a particular image I that has been segmented by an algorithm, let $C \in I$ be the segmentation of the cell corresponding to cell C_{GT} in the ground truth sementation of I, as described in the previous paragraph. A pixel $(x, y) \in C$ is called a false positive pixel if $(x, y) \notin C_{\mathrm{GT}}$. A pixel $(x, y) \notin C$, but $(x, y) \in C_{\mathrm{GT}}$ is labeled a false negative pixel. Let t denote the total number of pixels in C and C_{GT} and let t_{FP}, t_{FN} be the total number of false positive and false negative pixels, respectively.

The misclassification percentage, p_C associated with cell C is the value

$$p_C = 100 \times \frac{t_{\mathrm{FP}} + t_{\mathrm{FN}}}{t} \tag{1}$$

$0 \leq p_C \leq 100$. The data from the comparison experiment are these percentages p_C ranging over all the segmented cells. It is assumed that the experiment was performed so that this set of percentages $S = \{p_i, i = 1, \ldots, n\}$ is a random sample from a true, but unknown algorithm misclassification probability density $f(p)$, n being the total number of segmented cells. A histogram estimate of $f(p)$ based on S is defined as

$$\hat{f}_h(x) = \frac{1}{2hn}\,[\text{no. of } p_1, \ldots, p_n \text{ falling in } (x-h, x+h)] \tag{2}$$

and a continuous approximation to this histogram is the kernel density estimate

$$\hat{f}_K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - p_i}{h}\right) \quad \int_{-\infty}^{\infty} K(x)dx = 1 \tag{3}$$

where h denotes the bandwidth or smoothing parameter (see Silverman [19]). In our histograms, we used the Freedman–Diaconis choice of histogram bandwidth $h = 2\mathrm{IQR}/n^{1/3}$, where $IQR = Q_{0.75} - Q_{0.25}$ is the interquartile range and $Q_{0.75}, Q_{0.25}$ are the third and first quartiles, respectively. And, in our kernel estimate $h = 0.9\min[\hat{\sigma}, IQR/1.34)]n^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation of the misclassification data. The Gaussian kernel $K(x) = \exp(-x^2/2)/\sqrt{2\pi}$ is used. The kernel estimate converges to the true misclassification probabity density as n approaches infinity.

The CDF of $\hat{f}_K(p)$ is the continuous nondecreasing function defined as

$$F(p) = \int_0^p \hat{f}_K(x)dx \quad 0 \le p \le 100. \tag{4}$$

In this study, we used this continuous approximation to the empirical histogram

$$F_h(x) = \frac{1}{n}\{\#\, p_i \le x, i = 1, \ldots, n\}.$$

In our comparison of algorithms, their misclassification CDFs are compared. The optimal error free algorithm has CDF which is the unit step function. So, the closer a CDF is to the unit step function the better its accuracy with respect to ground truth. When comparing two algorithms A and B with misclassification CDFs, F_A and F_B, A is said to be more accurate at percentage point p if $F_A(p) > F_B(p)$. That is, the likelihood that the percentage of misclassified pixels will be p or fewer is larger for algorithm A than for B. If this happens uniformly over all p, then A unambiguously is the best algorithm by this criterion. If the majority of the cells have misclassification percentages less than p_0, then uniformity over $0 \le p \le 100$ may be replaced with uniformity over $0 \le p \le p_0$. Thus, those cells with $p > p_0$ are not considered in the CDF comparison. If uniformity does not hold for any reasonable p_0, then we rely on other population statistics to help make a decision, such as the shape and spread of probability density functions, sizes of means, and variances.

D. Shape Descriptors

To get a better understanding of how misclassification errors arise, three morphological descriptors: area, roundness, and roughness are investigated. The goal is to determine if an algorithm's segmentation misclassification errors are attributable either to inaccurate cell area, roundness, or roughness as compared to the ground truth cell area, roundness, and roughness. Inaccurate area usually means too many pixels or too few pixels were included in the segmentation of a cell and inacurrate roughness means too much or not enough smoothing of a cell boundary. Each of these characteristics have been shown to be important for assessing cell function and may be useful for benchmarking the phenotypic state of a cell culture.

The 2D projected size of the cell is taken as the area index, that is the area of the polygon formed from the calculated boundary points, rather than pixel area. Roundness describes how much the cell shape differs from a circle and the roughness index is the root-mean-squared average deviations between the calculated cell boundary coordinates and their center of mass. The roughness index increases as corners and edges are added. Roundness is defined as

$$\text{roundness} = 4\pi \frac{\text{area}}{\text{perimeter}^2} \quad 0 \le \text{roundness} \le 1. \tag{5}$$

That roundness is always less than one, with a circle having roundness one, follows from the isoperimetric inequality [14].

TABLE II
NUMBER OF GROUND TRUTH CELLS FRAGMENTED OUT OF 223 TOTAL A10 CELLS AND 401 NIH 3T3 CELLS

	(A10 Cells)			(NIH 3T3 Cells)		
	Long	Medium	Short	Long	Medium	Short
Otsu	25	28	27	22	38	42
K-Means	4	10	16	3	8	10
Canny	5	7	9	2	3	13
Watershed	2	5	7	1	2	3

For a cell with boundary pixels at $r_1, r_2, \ldots, r_n$, roughness is defined as

$$\text{roughness} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2} \quad \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{6}$$

where $d_i = \text{dist}(r_i, \bar{r})$ is the Euclidean distance between the center of mass $\bar{r}$ and the ith boundary pixel. For comparison, the roughness of a circle is zero. Roundness is unitless, where as roughness is in units of distance between pixels.

For each algorithm, area, roughness, and roundness indexes are computed. Histogram, kernel estimates and CDFs are formed for each of these indexes and compared.

An additional metric that counted the number of cells that were fragmented into more than one object was also collected during image processing and is summarized in Table II.

E. Manually Drawn Segmentations of Cells and Reference Data

To generate reference cell segmentation masks used in evaluating the algorithms, fluorescent images of the same 50 fields were also acquired with the CCD set to 1×1 binning under optimal filter conditions. The 1×1 binning provided an effective increase in resolution by using the minimum pixel size on the camera. Cell objects in these images were identified by manually segmenting cells or groups of touching cells using the ImageJ software package found at the NIH website. To facilitate the accurate identification of the cell boundary each cell was enlarged with the zoom tool. The paint brush tool was then used to outline each cell in the frame using a brush width of two pixels. Small processes that extend from the cells were only included if the edge of the process could be clearly visualized during the manual segmentation. Reference data from the expert drawn cell image masks were generated using the particle analyzer in ImageJ.

In order to validate our manual segmentations as ground truth, we had the person doing the manual segmentations make replicate segmentations on multiple cells. This gave us a measure of repeatability of the manual segmentations. Three replicate segmentations of a cell are shown in Fig. 2(a). The repeatability of our manual segmentation appears acceptable, since the shape and size features of the cell are approximately the same, except for deviations at local features. Also, the ground truth was validated by a second experiment. A second set of independent manual segmentations were performed under the supervision of an expert. Considering this segmentation as a fifth algorithm along with Canny, Otsu, K-means, and watershed algorithms, they were compared to our reference ground truth. The results

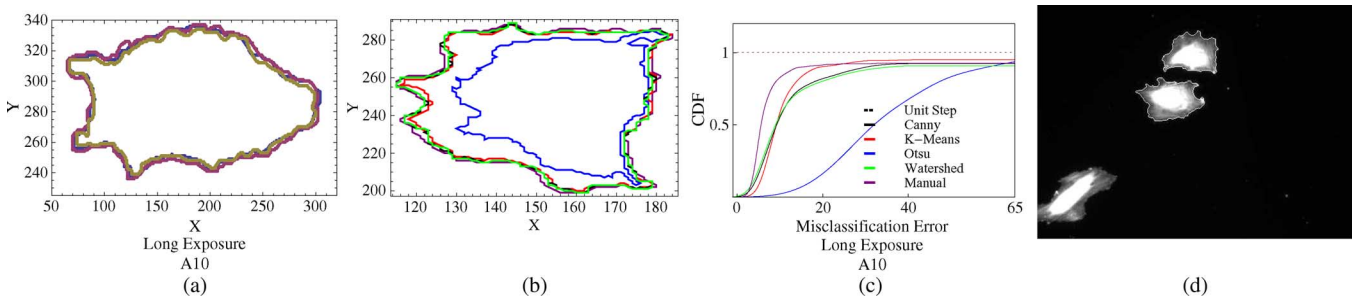

Fig. 2. (a) Three manual segmentation replicates. (b) Segementations of an A10 cell at long exposure [use legend in (c)]. (c) CDF of manual compared to the others and the unit step function. (d) An image with two cells segmented by Canny.

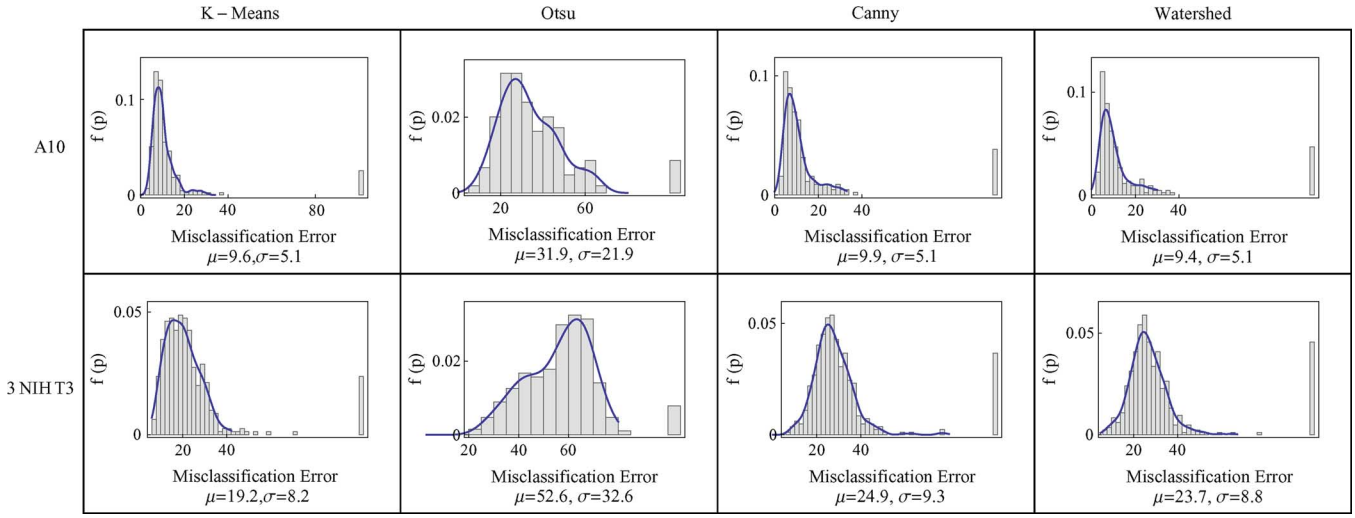

Fig. 3. Histogram and kernel (solid line) estimates of PDFs, $f(p)$ for NIH 3T3 fibroblasts and A10 cells at optimal (long exposure) setting.

of this comparison is shown in Fig. 2(c) for A10 cells at long exposure. The unit step function is shown as the optimal CDF. Fig. 2(c) shows that the misclassification CDF of the second segmentation is uniformly better than the misclassification CDFs of Canny, Otsu, K-means, and watershed. Because, the independent second manual segmentation is closer to the ground truth manual than any of the algorithms, the presumption is made that a manual segmentation accurately represents the boundary of the cell.

A visual comparison of a cell segmented manually and by the algorithms is shown in Fig. 2(b). Fig. 2(b) provides some clarification of where errors are made. The shape of the manually segmented cell is mostly preserved, but there are variations in area and boundary smoothness.

III. Results

A. Comparison at Optimal Setting

Plots in Fig. 3 of histogram and kernel estimates of misclassification percentages for A10 rat vascular smooth muscle cells and NIH 3T3 fibroblasts cells at long exposure times provide a summary of the algorithms' error distributions. These plots are based on a random sample of 223 Al0 cells from 50 images and 409 NIH 3T3 cells from 50 images. These sample sizes are large enough to provide accurate approximations to their

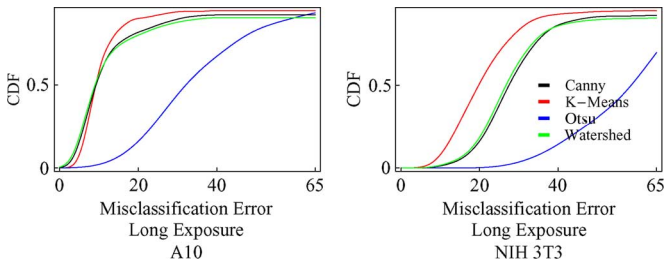

Fig. 4. Cumulative distributions, $F(p)$, for A10 and NIH 3T3 fibroblasts cells at optimal (long exposure) setting.

CDFs and PDFs. The plots illustrate that different morphologies between and within cell populations give rise to unique misclassification error distributions. The spikes at the far end of the right tail are artifacts that are explained in Section II-C. For A10 cells, the Otsu algorithm has a large misclassification error mean and standard deviation as well, and poorly approximates ground truth compared to the other algorithms. This may be because Otsu, a threshold algorithm, does not use as much information about the relationships between the pixels as the other algorithms. Canny, k-means and watershed statistics are comparable, but there are distinct differences. All their PDFs are right skewed, but k-means has a lighter right tail than the others. Their PDFs have a lognormal appearance. Their misclassification means and variances differ. For cells with morphologies

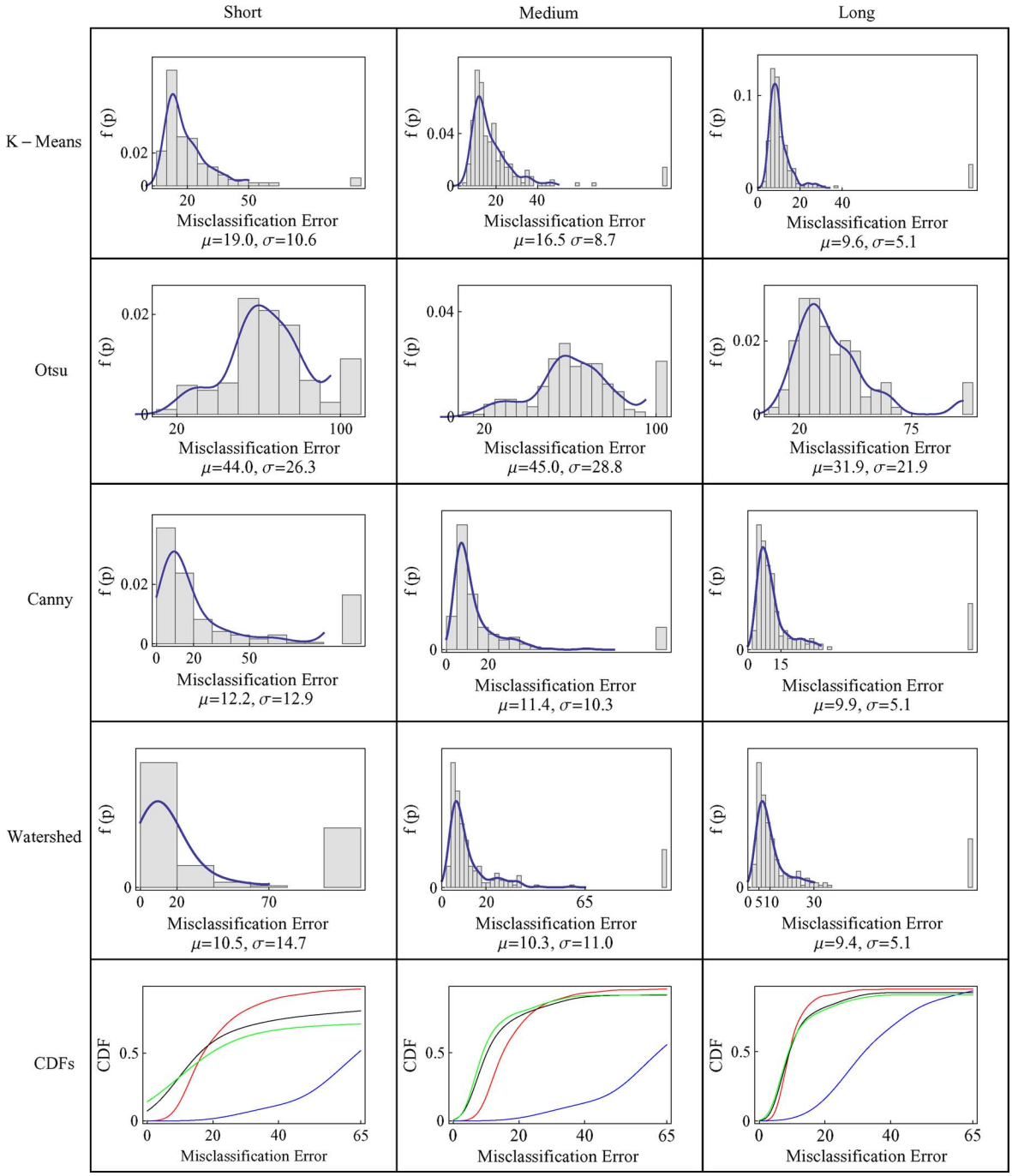

Fig. 5. A10: comparison of kernel estimates (solid line) at three exposure settings and CDFs: See Legend in Fig. 4.

similar to A10 cells, one can expect an error rate approximately 9.6% when using k-means, which is slightly smaller than Canny and watershed mean error rates.

The error distributions for the NIH 3T3 cells differ considerably from those for A10 cells. These algorithms when applied to cell populations with morphologies similar to NIH 3T3 cells are less accurate and more variable. Because the NIH 3T3 cells are considerably smaller than the A10 cells, their relative error means are about two times larger than the error means for A10 cells. The NIH 3T3 error distributions of Canny and watershed are almost symmetric about their means, while k-means is slightly right-skewed. A normal assumption for their error distributions may be appropriate. Otsu in this case is left skewed, an undesirable quality. The k-means error distribution has mean 19.2% and it has the smallest mean among the algorithms.

Fig. 4 shows a comparison of CDFs. The CDFs of Canny, k-means and watershed are almost indistinguishable for A10 cells with k-means slightly better. Their CDFs are uniformly higher than Otsu's. Ostu performs poorly as a segmentation algorithm for both NIH 3T3 and A10 type cells. For NIH 3T3 cells, k-means CDF is uniformly higher than the CDFs of the other algorithms over all $0 \leq p \leq 100$. Thus, it is the best algorithm at long exposure for NIH 3T3 cells.

One may conclude, based on the classification CDF comparison and the other information in the first two paragraphs of this section, the following. For cells similar to A10 cells, k-means is

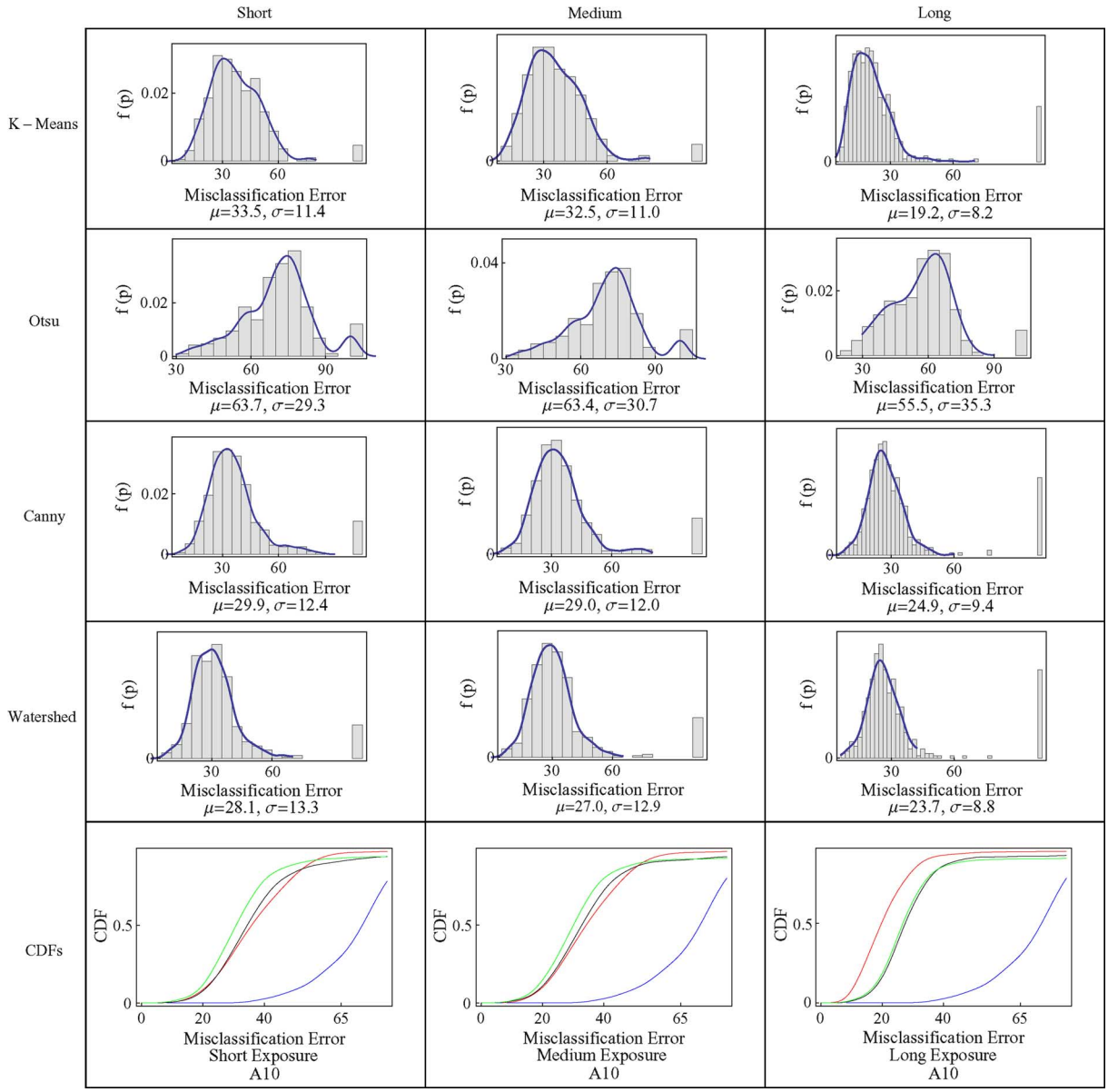

Fig. 6. NIH 3T3: comparison of kernel estimates (solid line) at three exposure settings and CDFs: See Legend in Fig. 4.

only marginally better than Canny and watershed at long exposure times. For NIH 3T3 cells, k-means is uniformly better than the other algorithms.

B. Robustness of Algorithms

In this section, we determine how varying exposure settings from short, medium to long affect the misclassification distributions of the algorithms, i.e., how robust are these algorithms over exposure settings. We determine if there are significant changes in their PDFs, CDFs, means, variances, and accuracies. The plots in Figs. 5 and 6 are based on the random samples used in Section III-A.

The signal-to-noise ratio increases when going from short to medium to long, see Table I, and as expected, the misclassification error means decrease as the signal-to-noise ratio increases for each algorithm. In all cases, except Otsu, there is a decrease in their error standard deviations, as well. For A10 cells, k-means seems to be affected most by exposure settings; its mean decreases by about 48% and its standard deviation decreases 52% in going from short exposure to long exposure. In going from short exposure to long exposure, Canny's mean and standard deviation change by 19% and 60%, respectively, and watershed's mean and standard deviation change by 11% and 65%, respectively. For NIH 3T3 cells, the following percent changes in means and standard deviations $\mu\%(\sigma\%)$ are observed to occur in going from short to long; k-means 63%(28%), Canny 17%(24%), and watershed 16%(34%). One can conclude that Canny and watershed are not significantly affected by a low signal-to-noise ratio, except by an increase in variability. Importantly, this suggests that these two algorithms are reasonably robust and can be used effectively over a variety of staining and acquisition settings.

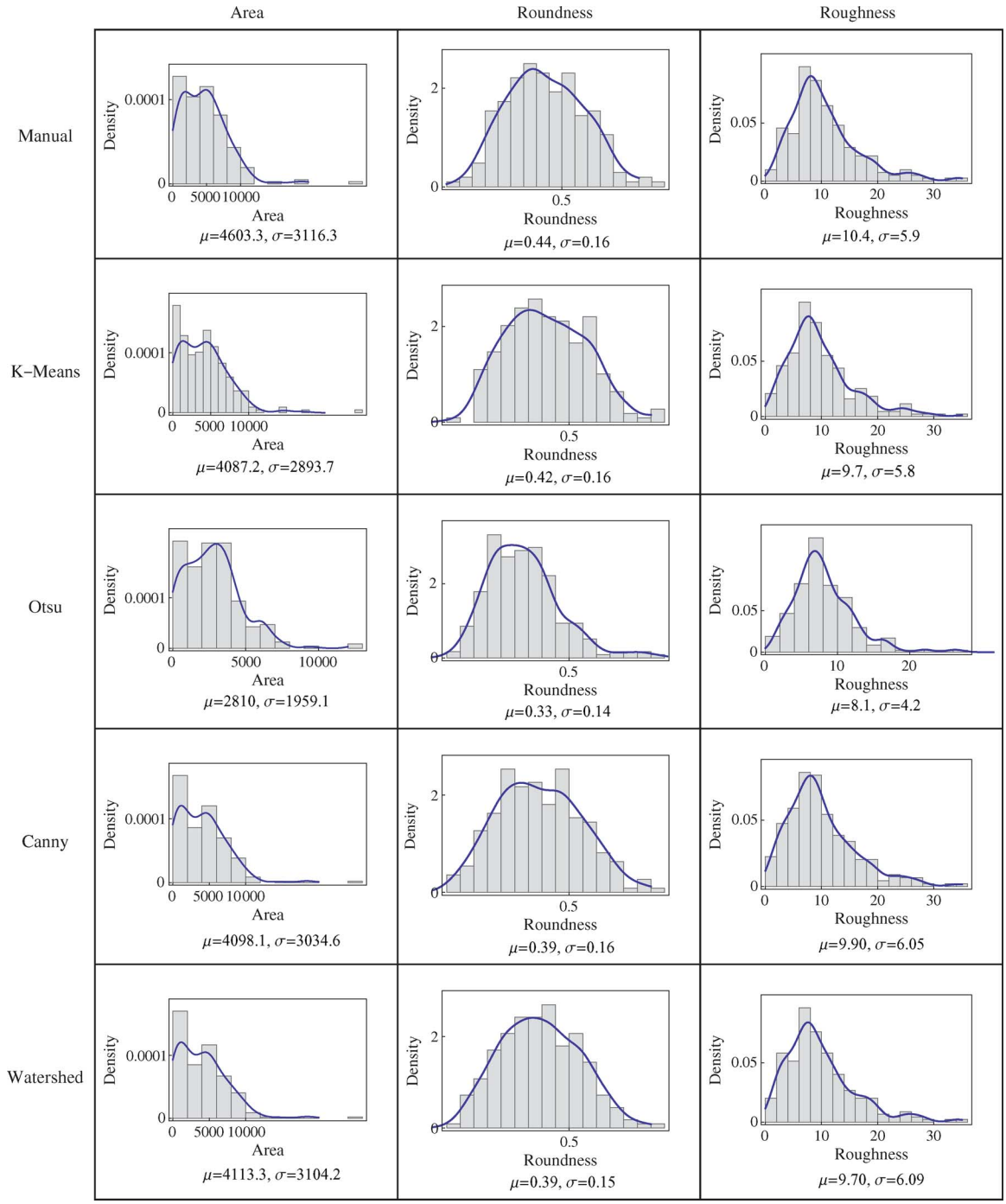

Fig. 7. A10: area, roundness, and roughness histograms and kernel densities (solid line) comparisons.

For A10 cells at short exposure using CDFs, one cannot definitely say either k-means, Canny, or watershed is the best algorithm. For A10 cells at medium exposure using CDFs with a threshold set at $p_0 \approx 25$, k-means is uniformly less accurate than Canny and watershed, with watershed slightly better than Canny. For CDFs associated with NIH 3T3 cells, the watershed algorithm is uniformly best for short and medium exposures at a threshold of approximately 60%. For long exposure k-means is uniformly the best over the entire range of percentages. The results show that the k-means algorithm improves fastest as the signal to SNR increases, but Canny and watershed are least sensitive to the SNR.

C. *Fragments*

Table II contains the fragmentation counts due to inadequate grouping of the classified pixels. To evaluate if misclassification results in significant cell pixel grouping errors (i.e., cell fragmentation), we counted the number of cell objects that were fragmented into more than one piece. This is shown in Table II.

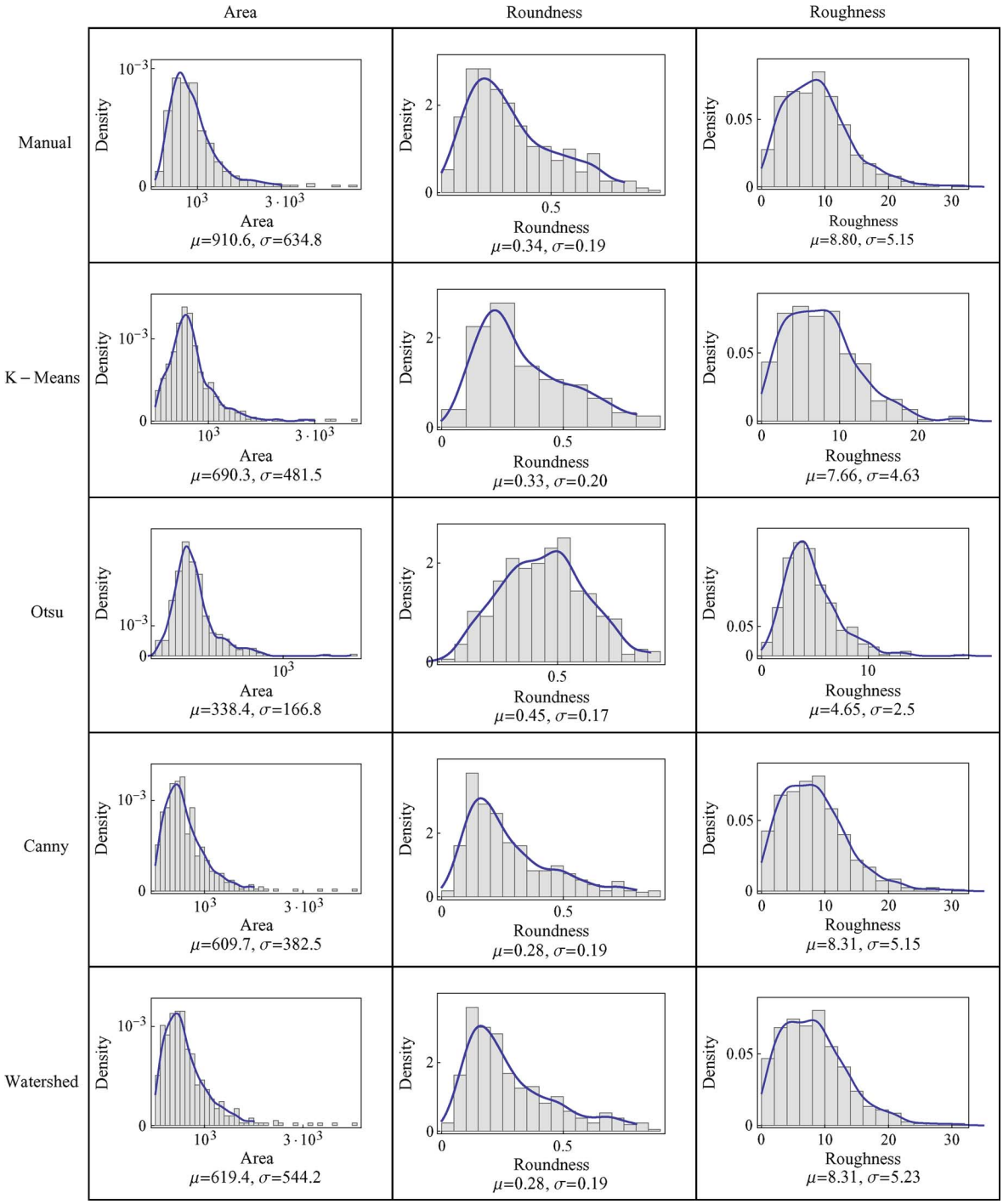

Fig. 8. NIH 3T3: area, roundness, and roughness histograms and kernel densities (solid line) comparisons.

As seen, fragmentation errors due to propagation of the misclassification errors to the grouping part of the segmentation algorithm are also dependent on signal to noise ratios. K-means performs well at high signal to noise ratios, but results in significant fragmentation of cell objects at lower ratios especially for the smaller NIH3T3 cells. The Canny algorithm also shows fragmentation errors that are sensitive to the signal-to-noise ratio. The watershed segmentation algorithm clearly resulted in the lowest number of fragmentation errors at all signal to noise ratios tested. Otsu performed the worst with the highest fragmentation error for all five conditions. Up to 5% and 20% of the cell objects in the NIH3T3 and A10 cell image series were fragmented with this segmentation algorithm, respectively.

D. Shape Descriptors

Using the same random samples of Section III-A, histogram plots of area, roundness, roughness indexes and their associated means and standard deviations for segmented A10 fibroblasts cells are shown in Fig. 7. Most noticeable in Fig. 7, Otsu has area, roundness and roughness much lower than the manu-

TABLE III
COMPARISON SUMMARY BASED ON CDFS

	Long	Medium	Short
A10	Canny, k-means, watershed all similar	Canny, watershed similar and better than k-means	watershed slightly the best
NIH 3T3	k-means the best	watershed slightly the best	watershed slightly the best

ally segmented cells and than the cells segmented by the other three algorithms. Otsu's errors originate from severely falsely labeling cell pixels as background, resulting in cells with reduced areas, overly smoothed boundaries and reduced roundness. As well, the other algorithms falsely labeled too many cell pixels as background (especially the small cells) producing areas on average smaller than the manually segmented cells. The k-means, Canny and watershed algorithms produce cells with similar roundness and roughness as the manually segmented cells. For the A10 cells, most of the misclassification errors can be attributed to falsely labeling cell pixels as background.

The plots in Fig. 8 are the shape descriptor distributions for NIH 3T3 cells. Here too, Otsu's errors result from severely labeling too many cell pixels as background, reducing roundness as compared to manual segmentation results and Otsu smooths the boundaries more than the manual segmentation. Also for this type of cell, k-means, Canny and watershed algorithms falsely label cell pixels as background, thus producing cells with smaller area than the manually segmented cells. They produce fewer large cells than the manual segmentations. K-means, Canny and watershed have similar roundness and roughness indexes as the manual segmentations. Since the right tails of their roughness PDFs are lighter than the manual's PDF, these algorithms produce fewer spiky cell boundaries than the manual.

IV. DISCUSSION

We proposed comparing cell segmentation algorithms using their misclassification error population statistics and by comparing fragmentation errors. In particular, we used misclassification error CDFs, PDFs, means and variances to compare Otsu, k-means, Canny, and watershed for segmentation of fluorescence images of fixed, stained cells. Manually segmented data were used as ground truth. The algorithm with uniformly largest CDF over the entire domain of the CDF is unambiguously the best classification algorithm. When there was no uniformly best CDF, other population statistics provided meaningful information for forming a conclusion. If the majority of the cells have fewer than p_0 of their pixels misclassified, then uniformity over $0 \leq p \leq p_0$ may be used as a replacement for uniformity over the entire range of the CDFs. This means that a small set of cells is not considered in the CDF comparison. Our evaluation process works when a large enough random sample of misclassification errors can be generated from the imaging experiment to accurately estimate the populations statistics. This strategy is different from a classic analysis of sensitivity and specificity in that only pixels in the union of the ground truth cell and the result of the segmentation algorithm are considered. There are no true negative pixels (background) in this union which is required to calculate a specificity value. We also evaluated fragmentation errors that can occur as a result of grouping the misclassified pixels. Both of these assessments are important, since depending on where misclassified pixels lie in the cell object, they can lead to inappropiate fragmentation of the cell object.

Two different cell populations having significantly different shapes, NIH 3T3 mouse fibroblasts cells and A10 rat vascular smooth muscle cells were imaged at short, medium and long exposures. Fifty images with approximately four or five cells per image with a total of 223 cells from a culture of A10 cells and 50 images containing approximately eight to nine cells per image with a total of 401 cells from a culture of NIH 3T3 cells were used in this study. The NIH 3T3 cells often have a small thin spindly appearance, while the A10 cells are large well spread cells that often have smooth edge appearances. Consequently, our results may be generalized, since the shapes of these two cell lines are representative of many cells used in cell biology and provide two contrasting morphologies. In real experiments, other parameters vary other than cell type and exposure settings, e.g., illumination level, filter type, and cell density. Ideally, an algorithm's performance should be fairly robust over all such parameters. Robustness over exposure settings was only considered in this analysis. Our image sets were collected over three variable exposure conditions (short, medium, long) to allow testing of the image signal-to-noise ratio on the performance of an algorithm. Table III contains what was concluded from our analysis of CDFs.

We found that all these algorithms perform better at a long exposure setting for both A10 and NIH 3T3 cells and are more accurate for the larger and rounder A10 cells, although Otsu's performance is much worse than the other three. Also, by comparing with ground truth the shape indexes, area, roundness and roughness of the algorithms, it was found that most of the errors are due to falsely labels cell pixels as background, producing cells with smaller area. In addition, Otsu oversmooths the boundaries and gives incorrect roundness values. We also found that watershed fragments less cells than the other algorithms.

Importantly, our results suggest that for quantitative characterization of a population of cells that have been experimentally seeded on a substrate at low density and stained with a high contrast edge stain, cells imaged at longer exposure times can be robustly segmented by many different algorithms giving similar misclassification rates and fragmentation errors. This information may be important for generating reference cell image data that describe a population of cells. Several algorithms can provide similar quantitative segmentation results from the images suggesting that high exposure is more robust to algorithm variations.

ACKNOWLEDGMENT

The manual segmentations were done by A. Kakkad, S. Jaffe, and J. Strand. The authors want to express their thanks to each of them.

REFERENCES

[1] V. Abraham, D. Taylor, and J. Hastings, "High content screening applied to large-scale cell biology," *Trends Biotechnol.*, vol. 22, no. 1, pp. 15–22, Jan. 2004.

[2] A. Carpenter, "Image-based chemical screening," *Nature Chem. Biol.*, vol. 3, no. 8, pp. 461–465, Aug. 2007.

[3] M. Halter, A. Plant, J. Tona, and J. Elliott, "Cell volume distributions reveal cell growth rates and division times," *J. Theoretical Biol.*, vol. 257, no. 1, pp. 124–130, 2009.

[4] J. Elliott, A. Tona, and A. Plant, "Comparison of reagents for shape analyis of fixed cells by automated fluorescence microscopy," *Cytometry*, vol. 52A, no. 2, pp. 90–100, 2003.

[5] M. James, *Classification Algorithms*. New York: Wiley, 1985.

[6] L. Coelho, A. Shariff, and R. Murphy, "Nuclear segmentation in microscope cell images: A hand segmented dataset and comparison of algorithms," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2009, pp. 518–521.

[7] P. Bamford, "Empirical comparison of cell segmentation algorithms using an annotated dataset," in *Proc. 2003 Int. Conf. Image Process.*, 2003, vol. 3, p. II-1073.

[8] D. Hodneland, N. V. Burkoreshtliev, T. Eichler, X. Tai, S. Furke, A. Lundervold, and H. Gerdes, "A unified framework for automated 3-d segmentation of surface-stained living cells and a comprehensive segmentation evaluation," *IEEE Trans. Med. Imag.*, vol. 28, no. 5, pp. 720–738, May 2009.

[9] E. Gelasca, B. Obara, K. D. Fedorov, Kvilekval, and B. Manjunath, "Biosegmentation benchmark for evaluation of bioimage analysis methods," *BMC Bioinformatics*, vol. 10, pp. 1–12, 2009.

[10] B. S. Database [Online]. Available: http//www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/, 2001

[11] C. Pantofaru and M. Hebert, A comparison of image segmentation algorithms Carnegie Mellon, Pittsburgh, PA, Tech. Rep. CMU-RI-TR 05-40 666/1999, 2005.

[12] A. Dima, J. Elliott, J. Filliben, M. Halter, A. Peskin, J. Bernal, M. Kociolek, M. Brady, H. Tang, and A. Plant, "Comparison of segmentation algorithms for fluorescence microscopy images of cells," *Cytometry*, vol. 79A, no. 7, pp. 545–559, Jul. 2011.

[13] , D. L. Taylor, J. R. Haskins, and K. G. , Eds., *Methods in Molecular Biology*. Totowa, NJ: Humana, 2006, vol. 356.

[14] M. P. do Carmo, *Differential Geometry of Curves and Surfaces*. Englewood Cliffs, NJ: Prentice Hall, 1976.

[15] K. Bowyer, C. Kranenburg, and S. Dougherty, "Edge detector evaluation using empirical roc curves," *Semiconductors*, vol. 84, pp. 77–103, 2001.

[16] J. Canny, "A computational approach to edge detection," *IEEE Trans. Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986.

[17] L. P. Coelho *et al.*, Nuclear segmentation in microscope cell images: A hand-segmented dataset and comparison of algorithms 2009 [Online]. Available: http://murphylab.web.cmu.edu/publications/157-coelho2009.pdf

[18] R. Gonzalez and R. Woods, *Digital Image Processing*. Upper Saddle River, NJ: Pearson Prentice Hall, 2008.

[19] D. J. Hand, *Construction and Assessment of Classification Rules*. Chichester, U.K.: Wiley, 1997.

[20] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probabil.*, 1967, vol. 1, pp. 281–297.

[21] T. Nattkemper, "Automatic segmentation of digital micrographs: A survey," in *MEDINFO*, M. Fieschi, Ed. *et al.*, Amsterdam, The Netherlands, 2004.

[22] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.

[23] Z. Perlman, M. Slack, Y. Mitchison, L. Wu, and S. Altschuler, "Multidimensional drug profiling by automated microscopy," *Science*, vol. 306, no. 5699, pp. 1194–1198, 2004.

[24] J. Russ, *The Image Processing Handbook, 4th Ed.*. Boca Raton, FL: CRC Press, 2002.

[25] D. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, pp. 605–610, 1979.

[26] Y. Zhang, "Evaluation and comparison of different segmentation algorithms," *Pattern Recognit. Lett.*, vol. 18, pp. 1335–1346, 1997.