16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

BENCHMARK SET OF SYNTHETIC IMAGES FOR VALIDATING CELL IMAGE ANALYSIS ALGORITHMS

Pekka Ruusuvuori, Antti Lehmussola, Jyrki Selinummi, Tiina Rajala, Heikki Huttunen, and Olli Yli-Harja

Department of Signal Processing, Tampere University of Technology
P.O. Box 553, 33101 Tampere, Finland
email: pekka.ruusuvuori@tut.fi

ABSTRACT

This article presents a synthetic image set for validation of cell image analysis algorithms. To address the problem of validation, we have previously developed a simulation framework for cell population images. Here, we apply the simulation for generating a benchmark set of cell images with varying characteristics. The value of simulation is in the ground truth information known for the generated images. Traditionally, the ground-truth has been obtained through tedious and error-prone manual segmentation of the images. While such approach cannot be fully replaced, we propose to use the simulated images for benchmarking along with manually labeled images, and present case studies of tuning and testing a cell image analysis algorithm based on simulated images.

1. INTRODUCTION

Cell image segmentation can be considered as a binary classification, where the pixels are classified as background and region of interest, i.e., cells. Inherently, such method (classifier) for performing the segmentation that would be globally optimal in different analysis tasks does not exist [1]. This fact leads to the current situation, where plenty of novel algorithms and variations of old ones are tuned for emerging applications. Algorithm development and tuning is needed especially in the field of biological or biomedical studies, where various kinds of cell populations are studied under different conditions, using different dye labels. While such tuning can in some cases be considered as normal engineering work, the validation of proposed algorithms still remains problematic, especially in the case of high-throughput measurements [2].

Manual analysis performed by an expert in the field of study is a commonplace as the basis of validation. The usual validation procedure includes a set of representative images that have been labeled by one or more experts. The labeled set is then used as a reference to which the results given by the developed analysis methods are compared. Given that the expert labeling has been performed rigorously, the procedure is valid but sometimes exhaustively laborious. In such cases where the decision requires special knowledge, expert analysis is the only way of validating the results.

Manual analysis, however, always includes uncertainty. Different people may have different criteria for making the decisions. For example in the case of cell images, one expert may label a bright spot as highly fluorescent cell, whereas another may consider it as noise or stain residue. Such discrepancies introduce between-analyst-variation to the results. In addition, bias between the analysts is possible, for example if different experts have different criteria for the minimum size of objects to be considered as cells. Another type of uncertainty is the within-analyst variation. Especially in routine tasks performed for a large number of samples, such as counting cells in images, the result given by a single analyst may vary. This kind of variation is even more obvious in laborious and non-trivial tasks, such as segmentation of cells, where the pixelwise result will almost certainly be different between trials made by the same person.

Automated image analysis has been presented as a solution for getting rid of the tedious task and the variation caused by analysts, see e.g., [3, 4, 5]. The automated image analysis algorithms are usually reproducible in the sense that they always produce the same output for the same input. They also treat all images in the same manner, without subjective bias or random errors. This does not mean that the result would be unbiased, but the possible bias will also be consistent. Importantly, the same facts concern also synthesis of ground-truth. When using simulated ground-truth images as reference, we exclude the possibility that the expert analyst would create bias or variance to the reference results. This helps in avoiding the kind of situation, where the analysis algorithm is tuned to follow the possible errors made by expert analyst into the reference results.

In our previous studies, we have introduced methods for simulating fluorescence microscopy images of cell populations [6, 7]. With the simulation platform [7], it is possible to generate arbitrarily large image sets with realistic characteristics. These images may be used for validation of image analysis algorithms, cell enumeration and segmentation in particular. Since the user has full control over the parameters, it is possible to generate images with varying properties, resembling for example population properties or growth. The benefit of parameter tuning is that it allows also incorporation of expert knowledge into the simulation [8].

In this article, we present a benchmark image set for validating cell image analysis algorithms. The purpose is to create a set that would be useful in developing and testing of novel cell image analysis algorithms. By providing a readily made set we provide possibility to compare their results with a common set, in other words to use the set as a benchmark. Moreover, since the image sets do not depend on a certain platform or require the use of SIMCEP simulator [7], the current study extends the group of potential users of the simulated images.

2. MANUAL APPROACH

The traditional approach for validating an image analysis algorithm is to use manual analysis. By analyzing manually a set of images, a ground truth is obtained. However, it is a well known fact that manual analysis produces variation to the results [9, 10, 2]. Thus, instead of a single ground truth, the results are actually a set of opinions. Here we present a case study, where three individual analysts outline cells from fluorescence microscopy images.

The images were acquired from human embryonic stem cell

This work was supported by the National Technology Agency of Finland and the Academy of Finland, project No. 213462 (Finnish Centre of Excellence program (2006 - 2011)), and partially supported by Tampere Graduate School in Information Science and Engineering (TISE) and Tampere University of Technology Graduate School.

derived neural precursor cell cultures. The fixed and immunohistochemically stained cells were mounted with Vectashield mounting medium containing DAPI (nuclear stain, Vector Laboratories) and imaged using fluorescence microscope (Olympus IX51S8F-2) equipped with a fluorescence unit and camera (Olympus DP71). Only the nuclear images of two separate sets, named here as image set 1 and 2, were taken further into the case study.

Three analysts performed manual segmentation for each of the images. In more detail, the perimeter of each nuclei was outlined by using the Adobe Photoshop software (Adobe Systems Inc, CA, USA). The result corresponds to the typical output of image segmentation, from which cell-level parameters related to the size and morphology can be calculated. Here, we calculated the compactness and solidity values for each object. Both compactness and solidity measure the circularity of the objects in slightly different manner. Compactness is defined as follows [11]:

$$Compactness = \frac{\sqrt{\frac{4}{\pi} Area}}{MaximumDiameter},$$

where the maximum diameter is obtained from an ellipse fitted into the object area. Solidity is defined as [11]

$$Solidity = \frac{Area}{ConvexArea},$$

where the convex area is given as the area of convex hull, that is, the smallest convex set containing the object pixels.

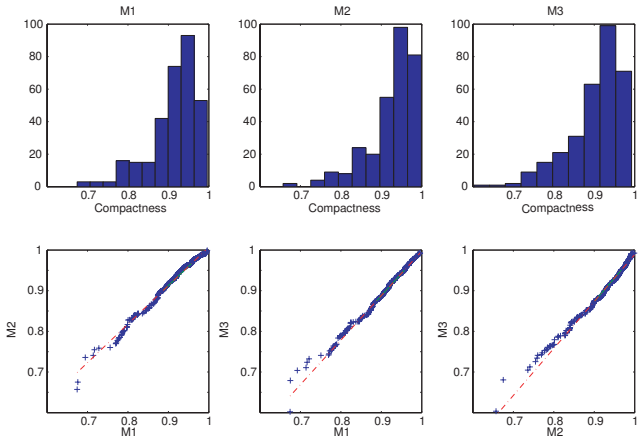

Fig. 1: Histograms of compactness measures calculated from image set 1 segmented manually by M1, M2 and M3, and quantile-quantile plots of the compactness values between the three analysts show differences between the results. A few clear outliers are left outside the plots.

The results for compactness measure for image set 1 given by three analysts M1, M2 and M3 are shown as histograms in Figure 1. In addition, the quantile-quantile plots between the three results are illustrated in Figure 1. Due to limited space, we exclude similar plots for the solidity measure and for image set 2. The graphs show that the results do not fully coincide; histograms between analyzers look different, and the quantile-quantile plots deviate from a straight line indicating that the distributions are dissimilar. Thus, we used Kolmogorov-Smirnov test [12] to find out if the result differ significantly. The results of the test are given in Table 1. For image set 1, both compactness and solidity results differ significantly in two out of three comparisons between analysts. For image set 2, the results between analysts M2 and M3 did not show significant difference for either measures, whereas the rest of the comparisons showed difference. Moreover, the estimates of the total number of cells in the six images vary between the analysts, but a rough compromise is around 800 cells. Here, only individual cells not touching to another are taken into the analysis. Thus, one source for the differences is that cells located close to each other can be marked as touching or completely separated. Obviously different people had different criteria for segmentation. These results show how significant the variation between the results given by different persons may be, and underlines the unreliability present in manually obtained reference.

Table 1: Results (p-values) of Kolmogorov-Smirnov test between the manual analysis results. The result pairs showing significant (0.01) difference are marked with boldface.

	Image set 1		Image set 2	
Sets	Compact.	Solid.	Compact.	Solid.
M1 M2	$\mathbf{3.8 \times 10^{-4}}$	$\mathbf{2.2 \times 10^{-6}}$	$\mathbf{0.004}$	$\mathbf{\sim 0}$
M1 M3	0.246	$\mathbf{1.2 \times 10^{-7}}$	$\mathbf{0.003}$	$\mathbf{\sim 0}$
M2 M3	$\mathbf{2.4 \times 10^{-6}}$	0.081	0.994	0.071

Finally, a point worth mentioning is the time needed for generating the manual segmentation results. It took hours of time to get the results by three persons, even though the number of images per person was only six, having altogether around 800 cells. Based on this case study on manual segmentation, we conclude that since manual analysis does not necessarily produce an unquestionable ground truth, we may as well use simulation to produce images with varying properties.

3. BENCHMARK DATA

Benchmark datasets are commonly used in pattern recognition, and increasingly also in image analysis, see [8] for a list of examples. The benefit of benchmark data is that one is able to compare results with those of other approaches. Another benefit, which we consider as the primary reason for generating a benchmark of synthetic cell population images is the ease of getting a labeled ground truth set that can be used for algorithm development and validation. It may be that the benchmark data does not exactly fit to the addressed research problem, but it may still provide beneficial information about the performance of the algorithm. We believe this is the case also with our cell images. In this Section we will describe essential parameters of the simulation process, and properties of the benchmark datasets.

3.1. Simulation

The proposed benchmark data set is generated using the previously introduced simulator for fluorescent cell populations [7], which is strongly based on parameterized random models for cell shapes. Other approaches for cell image synthesis can be found from, e.g., [8, 13, 14]. Although the detailed description of the simulation is published previously, we here describe the simulation parameters and methods mostly used for characterizing the benchmark images. However, for more comprehensive understanding of the simulation methodology, we recommend to see [7].

Similarly as the real cells, the simulated cells consist of different components, such as nuclei, cytoplasm, and intracellular objects. The features mainly characterizing these components are the shape and size. The shape is simulated using a random model, which is controlled with parameters defining the level of distortion and size of the generated shapes. First, a random polygon with k vertices and

scale r is generated as

$$x_i(\theta_i) = r[U(-\alpha,\alpha) + \cos(\theta_i + U(-\beta,\beta))]$$
$$y_i(\theta_i) = r[U(-\alpha,\alpha) + \sin(\theta_i + U(-\beta,\beta))], \quad (1)$$

for $i = 1,...,k$, where $\theta_i \in [0, 2\pi]$ is the polar angle, and $U(a,b)$ a uniform distribution on the interval $[a,b]$. Finally, the vertices are connected with spline interpolation. Now the parameters α and β control the randomness of the shape. With varying parameter values, a large scale of objects with very different shape characteristics can be generated.

When considering microscope images on a level of cell populations, the disposition of cells is a very fundamental feature. For example, the cells can be spatially situated very sparsely, or they can be overlapping with each other and form clusters, which is a more typical case. In simulation, each generated cell is either placed on the image uniformly or assigned into a specific cluster with probability p_c. The clustered cells are located around the cluster centers according to a normal distribution. Since especially the overlapping cells pose challenges for automated image analysis algorithms, the simulator provides a parameter for controlling this property in the simulated images. When considering the set of pixels R_i defined by a simulated cell, the relative amount of overlap L_{ij} caused by the region of pixels R_j of another cell can be measured by

$$L_{ij} = \frac{|R_i \cap R_j|}{|R_i|}, \quad i \neq j,$$

where the operator $|.|$ is the cardinality of a set, and $L_{ij} \in [0,1]$. Thereby, the maximum amount of allowed overlap can be controlled with parameter L (e.g., $L = 1$ overlapping is not limited, $L = 0$, no overlap is allowed).

Typically, microscope images suffer from errors and artifacts originating from the imaging system. For example, nonuniform illumination can significantly degrade the resulting image by generating a varying intensity to the image background. The effect of nonuniform illumination is simulated by adding a second degree polynomial on the image. The characteristics of the illumination are controlled with parameters defining the horizontal and vertical illumination centers, and the illumination energy E_m. Information about other measurement errors is available in the original publication [7].

3.2. Benchmark sets

The three benchmark sets of synthetic cell population images described in the following can be downloaded from: `http://www.cs.tut.fi/sgn/csb/simcep/benchmark`. All sets contain simulated images and corresponding ground truth images where objects are represented as binary masks. It is planned that in the future the variety of benchmark sets will be increased to cover also other interesting parameters. The reader is also encouraged to use the SIMCEP simulator; by tuning the parameters it is possible to generate data for specific research problems.

3.2.1. Clustering with increasing probability

The first image set consists of nuclei images with five different values of clustering probabilities. For each value of clustering probability we simulate 20 images, each of which contains 300 objects. See Table 2 for relevant parameter values of the first image set. The first parameter settings produce images with no overlap ($L = 0$) or clustering of cells ($p_c = 0$), see the leftmost image in Figure 2. The four other settings do not limit overlapping ($L = 1$), and introduce overlapping with increasing probability, which is clearly visible in Figure 2. The set provides test material for object segmentation and separation with varying level of difficulty.

Table 2: Set 1: clustering with increasing probability.

Parameter	Value
Images / parameter settings	20
Objects / image	300
Probability of clustering	0, 0.15, 0.30, 0.45, 0.6
Background energy	0.25
Autofluorescence energy	0.25
Overlap limit L	0, 1, 1, 1, 1

3.2.2. Cells with nuclei, cytoplasm and subcellular objects

The second benchmark set consists of multichannel images, where nuclei, cytoplasm, and subcellular components have each been labeled into their own channels. Nuclei usually appear as rather compact, roundish objects, whereas the cytoplasm shape is more irregular. However, both nuclei and cytoplasm, as well as the small subcellular objects can be simulated with the same shape model given in Equation 1 by tuning the parameters such that size and randomness of the shape increase when cytoplasm is generated. The most important parameters are given in Table 3.

The set has three-channel images of two conditions, one with good quality and one with lower quality, meaning that overlapping and noisy background are introduced. Figure 3 illustrates one image from both conditions. The images can be used either as three-channel images by using all channels, or channel by channel for separate analysis of nuclei, cytoplasm, and subcellular objects. Nuclei, cytoplasm, and subcellular components all have their own binary mask for ground truth.

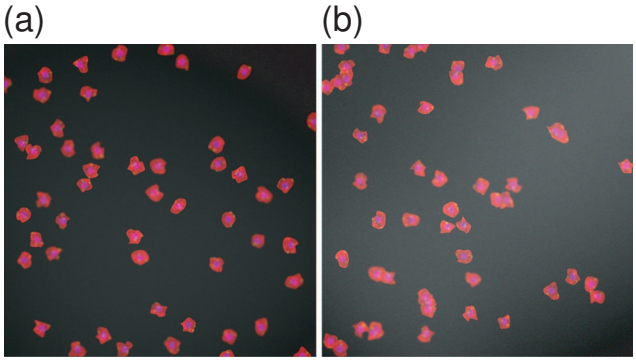

Fig. 3: Multichannel images with nuclei, cytoplasm, and subcellular components each stained for different channels without disturbing background (a) and with background illumination and slight overlapping introduced (b).

3.2.3. Cells from two populations

The purpose of this set is to serve as a test case for class discrimination. The two populations have slightly different characteristics, one having rather regular and round shapes and the other having more irregular shapes with more variation. Altogether 20 images of each population are available, which could, e.g., be used in training. Besides class determination, these images are potentially useful for testing feature selection and accuracy of segmentation based on desired features. Example images from both populations are shown in Figure 4. This time the most crucial parameters are related to shape and size, as listed in Table 4.

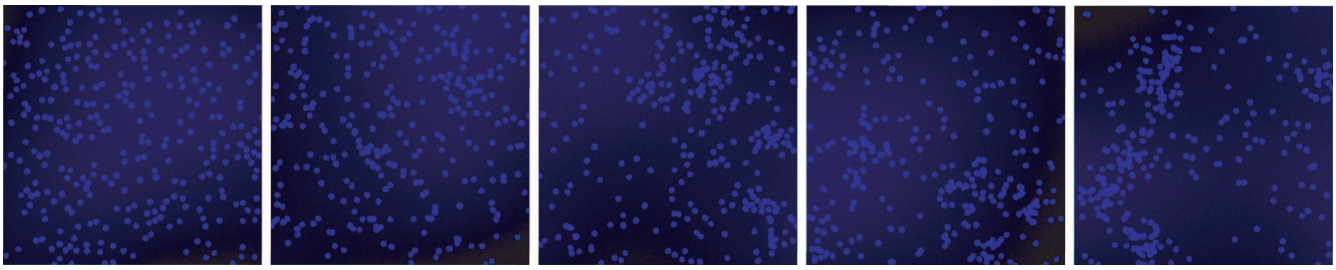

Fig. 2: Example images from sets with five different overlapping probabilities. The probability of clustering increases from left to right such that in the first image the probability is zero and in the last image the probability is 0.60.

Table 3: Set 2: Cells with nuclei, cytoplasm, and subcellular objects with (a) no overlap, and (b) slight overlap and disturbing background.

Parameter	Value
Images / parameter settings	20
Cells / image	40
Subcellular objects / cell	3
Probability of clustering	(a) 0 (b) 0.1
Background energy	(a) 0.15 (b) 1.0
Autofluorescence energy	(a) 0.05 (b) 0.25
Overlap limit L	(a) 0 (b) 1

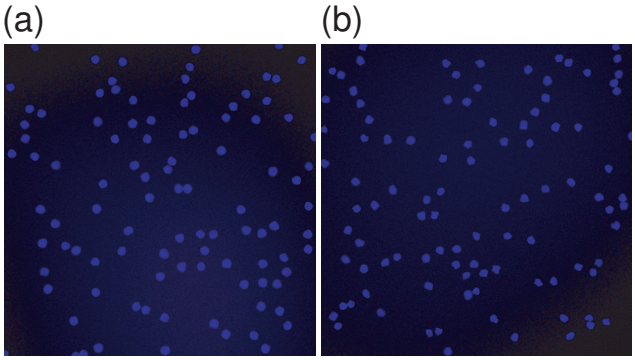

Fig. 4: Two populations with slightly different characteristics. (a) Objects are fairly round and symmetric. (b) Objects are slightly smaller, and the boundary is more irregular.

4. EXPERIMENTAL RESULTS

The comparison of the segmentation result and generated ground truth can be done in pixel-level, for example by using methods for quantifying discrepancy between segmentation result and binary ground truth mask presented in [15]. Here we show how the database images can be used for tuning the performance of CellC enumeration software [4]. CellC is a software tool designed for segmentation and object counting from fluorescence images of microbial populations. The current version CellC 1.2 enables also extraction of a few shape features. The first database set from Section 3.2.1 with 20×5 nuclei images, each having 300 objects, was used for the experiment. First, the parameters of CellC were slightly changed from the default settings in order to obtain a satisfactory result for the images with clustering probability set to zero. This situation resembles a case where algorithm development is done with images having well separable objects. As a result (black line in Figure 5 (a)), practically perfect results were obtained for the set with no clustering and overlapping. By running the analysis with similar settings for the rest of the set with increasing probability for clustering and overlapping allowed, severely degraded results were obtained. This is due to too conservative settings for cutting cells with suspicious shape, leaving substantial amount of ambiguous objects covering more than one ground truth object in the result.

Table 4: Set 3: Cells from two populations; (a) compact, roundish cells, and (b) more irregularly shaped cells.

Parameter	Value
Images / parameter settings	20
Cells / image	100
Radius	(a) 12 (b) 11
Shape parameters α and β	(a) 0.1, 0.1 (b) 0.2, 0.2

In images where overlap is allowed ($p_c > 0$), part of the cells are so heavily overlapping that automated separation is almost impossible. Moreover, tuning the parameters of CellC such that almost all suspicious cells would be split will lead to oversegmentation of normal cell shapes also. This would lead to intolerably large amount of false objects, when cells are split into parts. Thus, instead of optimizing the number of detected objects, we aimed at finding a suitable trade-off between under and oversegmentation. The result of tuning the performance of CellC for the sets with clustered and overlapping cells can be seen as the gray line in Figure 5 (a). The cell enumeration results presented here raise the question about true detection accuracy. Here we assume that the number of cells is the quantity of interest. However, the cases of under and oversegmentation should be taken into account such that, for example, a falsely split object would not compensate for a missed object. This would require a quantitative measure taking into account different error cases, which we will leave as the topic of another study.

Another case study demonstrates how the images in the third database set from Section 3.2.3 can be used for testing feature extraction. The images of two populations with slightly different characteristics were analyzed with CellC software. The area, length, and solidity features were chosen as the output, and the extracted features are visualized as boxplots in Figure 5 (b). The outlier values are not shown in the boxplot visualization. The results show that the segmentation has been accurate enough for revealing differences between populations. For example, the changes in the simulation parameter values (radius, shape) between the two synthetic populations have lead to differences in the measured size, length, and solidity. Furthermore, the boxplots for solidity feature suggest that other features would probably be needed, if the aim would be to discriminate objects based on shape descriptors. The database images could be used for testing and validation, for example, when developing new features for discriminating the differences between smooth

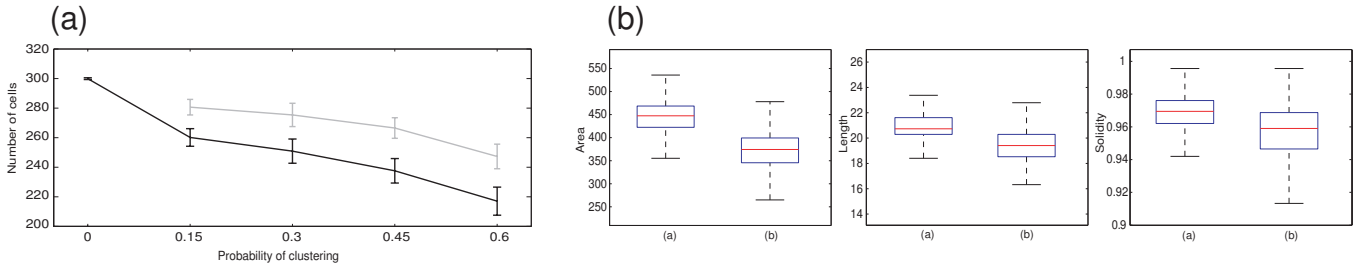

Fig. 5: (a) Tuning the performance of the CellC cell enumeration software with the database images. Each image has 300 objects. Black line shows the result when performance is tuned for well separable cells ($L = 0$), and gray line shows the result for sets with clustered and overlapping objects when parameter has been tuned with images having overlapping cells ($L = 1$, clustering probability p_c varies from 0.15 to 0.60). (b) Features for two populations extracted by CellC illustrated as boxplots. There were 20 images for each population. Outlier values are not shown in boxplots.

and slightly more irregular objects.

5. CONCLUSIONS

In this article, we presented a benchmark image set of simulated fluorescence microscopy images. By providing the image sets we offer a platform independent, easily accessible way for obtaining and using simulated cell population images. The benchmark data can be used for validation of, e.g., cell counting and feature extraction algorithms. The use of synthetic images was motivated by showing how manual segmentation may produce potentially inaccurate ground-truth data. The presented case studies demonstrated the usefulness of benchmark data in testing cell counting and feature extraction. The provided benchmark sets serve as a starting point for validation of analysis algorithms, but as the simulation framework is developed further, more sets can be added. In addition, the simulator is freely available and can be used for generating data sets that support other research topics.

Simulation of complex objects, such as cells, is a very challenging task. Since natural variation can not be fully included in a model, there will always be a limit in how natural the result will be. We have shown that in some cases simulation may provide valuable information that would be very hard to obtain with traditional approach. However, we are not suggesting that manual validation should be totally replaced by simulated images, but quite the opposite: there should also be more publicly available databases for manually segmented cell images. In the future, the benchmark data could be used for, e.g., comparing the accuracy of manual segmentation and automated analysis against the simulated ground truth.

6. ACKNOWLEDGMENTS

The authors would like to thank REGEA Institute for Regenerative Medicine, University of Tampere for providing the stem cell culture images, and Mr. Sharif Chowdhury for his help with programming.

7. REFERENCES

[1] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer Verlag, New York, 1997.

[2] X. Zhou and S. Wong, "Informatics challenges of high-throughput microscopy," *IEEE Signal Processing Magazine*, vol. 23, pp. 63–72, 2006.

[3] M.H.F. Wilkinson, *Digital Image Analysis of Microbes: Imaging, Morphometry, Fluorometry and Motility Techniques and Applications*, Wiley, 1998.

[4] J. Selinummi, J. Seppälä, O. Yli-Harja, and J.A. Puhakka, "Software for quantification of labeled bacteria from digital microscope images by automated image analysis," *Biotechniques*, vol. 39, pp. 859–863, 2005.

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

[6] A. Lehmussola, J. Selinummi, P. Ruusuvuori, A. Niemistö, and O. Yli-Harja, "Simulating fluorescent microscope images of cell populations," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, 01-04 Sept. 2005, pp. 3153–3156.

[7] A. Lehmussola, P. Ruusuvuori, J. Selinummi, H. Huttunen, and O. Yli-Harja, "Computational framework for simulating fluorescence microscope images with cell populations," *IEEE Transactions on Medical Imaging*, vol. 26, no. 7, pp. 1010–1016, 2007.

[8] T.W. Nattkemper, A. Saalbach, and T. Twellmann, "Evaluation of multiparameter micrograph analysis with synthetical benchmark images," in *Proc. 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, A. Saalbach, Ed., 2003, vol. 1, pp. 667–670 Vol.1.

[9] D. Webb, M.A. Hamilton, G.J. Harkin, S. Lawrence, A.K. Camper, and Z. Lewandowski, "Assessing technician effects when extracting quantities from microscope images," *J Microbiol Methods*, vol. 53, pp. 97–106, 2003.

[10] T.W. Nattkemper, T. Twellmann, H. Ritter, and W. Schubert, "Human vs machine: evaluation of fluorescence micrographs.," *Comput Biol Med*, vol. 33, no. 1, pp. 31–43, Jan 2003.

[11] J.C. Russ, *The Image Processing Handbook*, CRC Press, Boca Raton, USA, 3rd edition, 2000.

[12] F.J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[13] T. Zhao and R.F. Murphy, "Automated learning of generative models for subcellular location: building blocks for systems biology.," *Cytometry A*, vol. 71, no. 12, pp. 978–990, Dec 2007.

[14] D. Svoboda, M. Kašík, M. Maška, J. Hubený, S. Stejskal, and M. Zimmermann, "On simulating 3D fluorescent microscope images," in *Computer Analysis of Images and Patterns, Berlin, Heidelberg: Springer-Verlag*, 2007, pp. 309–316.

[15] Y.J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, 1996.