# Automatic Analysis of Leishmania Infected Microscopy Images via Gaussian Mixture Models

**Pedro A. Nogueira, Luís Filipe Teófilo**

Laboratória de Inteligência Artificial e Ciência de Computadores
Faculdade de Engenharia da Universidade do Porto, FEUP
Porto, Portugal
pedro.alves.nogueira@fe.up.pt

**Abstract.** This work addresses the issue of automatic organic component detection and segmentation in confocal microscopy images. The proposed method performs cellular/parasitic identification through adaptive segmentation using a two-level Otsu's Method. Segmented regions are divided using a rule-based classifier modeled on a decreasing harmonic function and a Support Vector Machine trained with features extracted from several Gaussian mixture models of the segmented regions. Results indicate the proposed method is able to count cells and parasites with accuracies above 90%, as well as perform individual cell/parasite detection in multiple nucleic regions with approximately 85% accuracy. Runtime measures indicate the proposed method is also adequate for real-time usage.

## 1 Introduction

Leishmania is the parasite responsible for Leishmaniasis, a disease currently affecting over 12 million people throughout 88 countries [1]. Leishmaniasis is treatable by chemotherapeutics, which, nevertheless, suffer from poor administration regimens and high host toxicity [2]. Although the disease is not generally deadly, it severely damages the immune system, leaving the body exposed to other deadly pathogens, which often prove fatal [3]. The inadequate means to treat Leishmaniasis render the research for new treatments an urgent task.

Research in microscopy imaging produces large amounts of data, which requires anywhere from full days to weeks to classify and annotate. In a single laboratory the number can easily ascend to thousands of images with merely a dozen different experiments. Not only does this detract the researchers from exploring new alternatives, as

adfa, p. 1, 2011.

it also introduces inter-person variance, as many images are extremely cluttered and contain several hundreds of cells and parasites. This results in a time consuming and mentally straining process, which expresses itself as a decaying function over time as the subject gets tired, frustrated or bored.

These reasons justify the need for the development of automatic mechanisms that are able to replace or aid researchers in the annotation task, for which and to the best of our knowledge no current solution exists. The proposed method provides a fully automatic, real-time pipeline for the identification of cells and parasites in Leishmania infected microscopy imaging, thus enabling more accurate annotations.

Pertaining this paper's organization, it is structured as follows. Section 2 describes the main characteristics of fluorescence microscopy imaging, as well as the dataset used in this study. Section 3 discusses the state of the art in cell identification and segmentation in microscopy imaging. Section 4 briefly describes the proposed method, followed by the description of its steps. In section 5, the results for the segmentation step and the implemented classifiers are presented. Finally, section 6, presents conclusions on the developed work, commenting on its performance and readiness for real-world applicability.

## 2 Fluorescence Microscopy Imaging

In contrast to the classical optical microscopy, the use of fluorescence microscopy allows simultaneous labeling of different cell components, which can be easily distinguished based on the fluorescence properties of their specific dyes [4]. The images collected for this study used three fluorophores, which emitted three distinct wavelengths [4]. These corresponded to the cell nuclei DNA (in blue), cytoplasmic and nuclear DNA (in red) and the parasitic DNA (in green). This provided three separate sets of data per image (Figure 1), motivating the identification of cells, parasites and cytoplasm individually in the three channels as independent images.

Although very popular, fluorescence microscopy imaging (FMI) presents some well-known issues that also characterized our dataset. The most noticeable issues included: non-linear illumination (due to poor lighting conditions and sub-optimal experimental setup), photo bleaching [5], varying contrast, Gaussian noise, chromatic aberrations and overlapping cells, as well as parasites (due to various focal planes).

In this study 794 fluorescence microscopy images from random drug trials with different experimental setups were collected and used. These images were collected through a light microscope and annotated manually by a Leishmania research team at the INEB/IBMC laboratory. Refer to section 6 for further details.

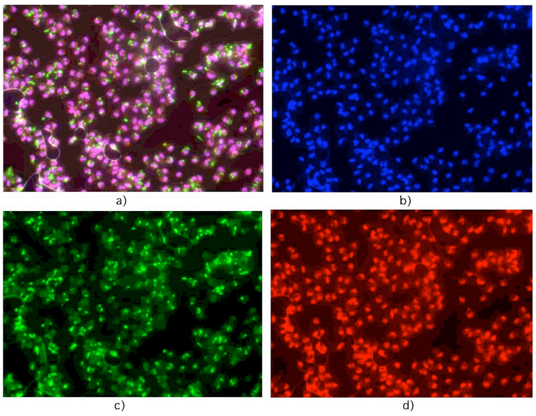

**Fig 1.** Details of a fluorescence microscopy image. a) Original image; b) Cell nuclei channel; c) Parasite nuclei channel; d) Cytoplasmic channel.

## 3 Related Work

Microscopy image analysis has been an active field for several decades. In related work, Liao et al. [6] used a simple thresholding method, coupled with mathematical morphology and contextual shape detection to detect white blood cells. However, their approach does not tolerate cells outside the defined conditions (e.g.: poorly segmented regions, forming a cell cluster region). An automated method for cellular membrane segmentation is described in [7]. This method also allows the reconstruction of unstained tracts through the nuclear membranes as a spatial reference. Jiang et al. have also proposed white blood cell segmentation using scale-space filtering and watershed clustering in HSV color space [8].

Park has proposed bone marrow cell segmentation through an iteratively relaxed watershed algorithm [9]. However, this work is sensitive to illumination and noise conditions, since it overly relies on the fixed mean color values of each patch for the relaxation procedure. Begelman [10] performs cell nuclei segmentation using color, shape features and a fuzzy logic engine. This work is more robust than the aforementioned one because the extracted shape features, serve as an auxiliary classification input. However, it is still not able to account for non-circular geometries or abnormally colored cells due to the implemented rules' simplicity.

Yu proposes using an adaptive thresholding technique to detect cell nuclei, which are then expanded via level sets to determine cell boundaries [11]. Yan proposes a similar approach [12]. Yan improves on Yu by replacing the adaptive thresholding with a distance map of the initial adaptive histogram-based thresholding step. This distance map is then used to create a watershed transform, serving as a region list representing the level-set seed points (Figure 2). The only drawback to this approach is that it is not able to deal with highly cluttered images, as the distance map would not provide enough information to accurately parameterize the watershed transform, thus leading to an erroneous number/location of seed points for the level-set step.

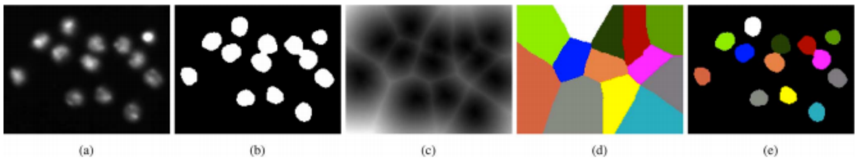

**Figure 2.** Segmentation example of the method proposed in [12]. (a) Patch taken from original DNA image. (b) Binary thresholding result of (a). (c) Distance transform of (b); (d) Result of the watershed algorithm on (c). (e) Labeling nuclei by combining (b) and (d).

From this review, it is clear that most of the literature does not attempt or is unable to deal with highly cluttered or overlapped image regions. This is a major concern in microscopy image analysis as the great majority of real-world data is heavily cluttered and saturated. The proposed method in this paper aims at addressing this issue.

## 4 Proposed Method

The proposed method focuses on developing robust methods for identifying and segmenting cellular and parasitic agglomerations in confocal microscopy images. The method starts by splitting the original image $f$ in three channels: $f_c$ (blue); $f_p$ (green); $f_{cyt}$ (red). Each channel is then normalized and segmented through Otsu's Method [13]. This yields three region vectors, corresponding each one to cellular DNA, parasitic DNA and weak[1] DNA signatures (cytoplasm). Low-level features are computed for the first two aforementioned region vectors, and then used to train a rule-based classifier and a Support Vector Machine, both of which attempt to classify each region as to how many cells or parasites it contains. To resolve disputes between these two classifiers, a voting system taking into account both of the classifiers' error margins is employed. Each region is then further segmented into the predicted number of sub-regions by Gaussian unmixing [17]. Figure 3 can be inspected bellow for a more structured understanding of the described pipeline.

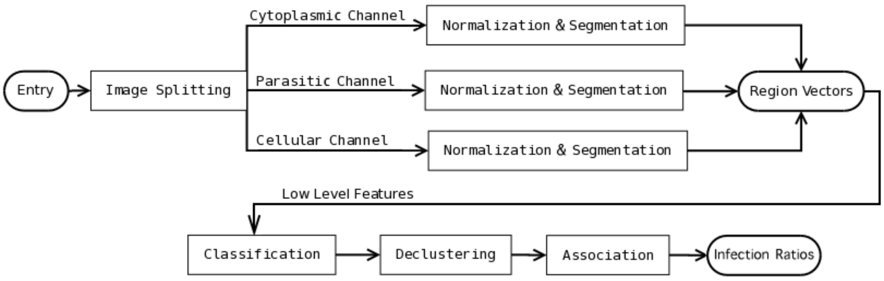

**Figure 3.** Developed method's architecture.

[1] In order to register a weak DNA signature, this fluorophore must be highly sensitive. Thus, it also registers the cell nuclei's DNA. Since the cell nuclei can be trivially subtracted through set operations involving the cell channel, we denominated this channel as the cytoplasmic channel.

### 4.1 Pre-Processing and Segmentation

The method first splits each of the target image's f color channels, as they are independently processed. Each image channel is then normalized and segmented into background and foreground components. An initial study on the general image characteristics was conducted in order to choose an appropriate segmentation technique. In this study, the intensity values of 120 randomly selected images presented clear bimodal distributions for all color channels (Figure 4).

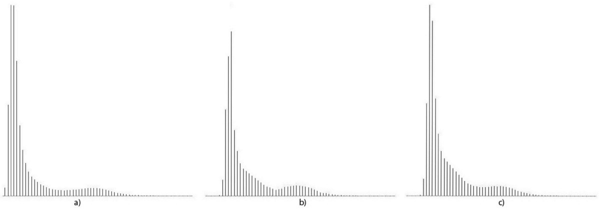

**Figure 4.** Bimodal distributions observed in the RGB components of 120 images (averaged). A) Red color component; b) Green color component; c) Blue color component.

Thus, Otsu's Method presented itself as a fitting approach due to its low temporal and spatial complexity, non-parameterisable characteristics and segmentation principle. Otsu's Method's principle assumes a bi-modal distribution in the target dataset, for which it attempts to determine an optimal threshold value $t$ by minimizing intra-class variance $(\sigma)$ [13]:

$$\sigma_b^2(t) = \sigma^2 - \sigma_\omega^2(t) = \omega_1(t)\ \omega_2(t)\ [\mu_1(t) - \mu_2(t)]^2 \qquad (1)$$

Each color channel was binarised using Otsu's Method and then proceeded to a connected component analysis [18], resulting in a region vector representing the cellular, parasitic and cytoplasmic regions present in the image. Note that the cytoplasmic regions are not used in this work, as they are intended for associating cell-parasite pairs. They were, however, computed and integrated into the method in hindsight of future work. Figure 5 depicts the result of the segmentation step for the cellular channel.

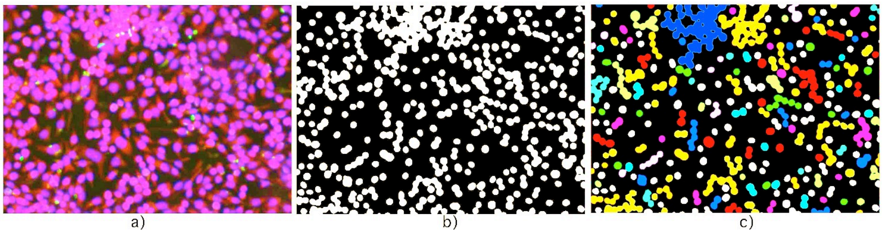

**Figure 5.** Segmentation output of a moderately cluttered image. a) Original image; b) Segmentation output; c) Visual representation of the cellular region vector obtained from the connected component analysis (randomly color-coded).

Following the region extraction, various low-level features were computed for each region. From these we highlight the ones with greatest classification potential, which were later chosen to form the following feature vector $F_i$=[$a$, $ll_{1..N}$, $d^1{}_{1..N}(ll)$, $d^2{}_{1..N}\ (ll)$]. The features comprising the vector are: $a$, the area value (in pixels); $ll$, the log-likelihood ratios for modeling the region with 1 to $N$ Gaussian mixtures; $d^1(ll)$, the first discrete derivate of $ll$ and $d^2(ll)$, the second discrete derivate of $ll$. The area feature was used to define the rule-based classifier. The log-likelihood ratios and their derivates were used to train the SVM classifier.

### 4.2 Classification

The classification step is based on the assumption that the regions obtained from the segmentation process may not always correspond to a single cell or parasite. Based on this, it attempts to discern in how many sub-regions each region must be split. This is achieved by employing two separate classifiers: a rule-based classifier (RBC) and a support vector machine (SVM) [20].

The RBC exploits the low area overlap percentage observed between cell and parasite pairs. This low overlap percentage is due to the high depth of field observed in the collected images, resulting in a near-perfect 2D cross-section of the 3D space within the tissue sample. Following this principle, and as single cells/parasites presented normal distributions, it was hypothesized that the area functions for cells and parasites could be approximated by a harmonic function. In fact, this was verified experimentally with the addition of larger multi-nucleic regions being less frequent. Thus, the functions present decreasing harmonic patterns. Although simple, this approach was found to be quite accurate (around 89.0% accuracy, hitting a maximum value of 98,9% when considering an error of ±2 regions as acceptable). Figure 6 illustrates a detail for one of the observed decreasing harmonic functions.

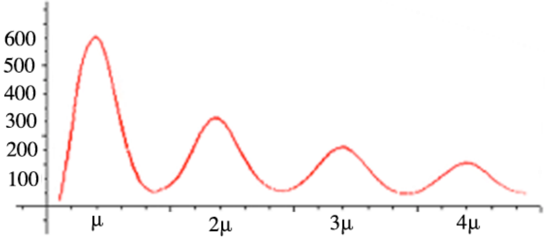

**Figure 6.** Detail of a decreasing harmonic function described by the normalized (cellular) area values of 120 randomly sampled images (~870 regions per image: 106.318 region samples total). Horizontal axis: area values; Vertical axis: area value occurrences.

In light of this finding, the classifier was programmed with rules reflecting these functions and taught to ignore values outside its knowledge space, as there was no data to accurately model regions over 10 cells or parasites. Note that cells and parasites each have their own harmonic function, since their area distributions do not exhibit the same standard deviation.

The SVM classifier relies on the concept that circles and ellipses can be described as Gaussian distributions and, as macrophages and parasites partake such geometry, clusters of these objects can be formulated as a mixture problem. Our conjectured hypothesis was that: as mixtures are added to the modeling process of each region, the improvement rate is described in the log-likelihood ratio evolution sequence. Thus, if an initial annotated dataset with the correct number of mixtures $N$ is available, it should be possible to model a function that is able to predict this *N* for new, non-annotated observations. Following this hypothesis, the classifier was trained with a subset of features from the main feature vector, consisting of the log-likelihood ratios, it's first and second order discrete derivates. The training set was obtained from roughly 150 regions, modeled with *N* = [1..15] Gaussian mixtures. Various machine-learning classifiers were tested. Ultimately, a SVM model was chosen as it achieved the highest (85,3%) sequential split and cross-validation classification rates.

### 4.3 Declustering & Association

Having obtained a prediction for the number of nuclei in each region, these were unmixed using the Expectation-Maximization [14] method. The algorithm was parameterized with a minimum standard deviation of $1x10^{-6}$, and a maximum of 200 iterations. To minimize runtime, the seeds for the centroids of each mixture were set by the averaged centroids of a 10 fold cross-validated K-Means. Figure 7, portrays the meth-od's expert performance, even in the presence of large nucleic clusters.

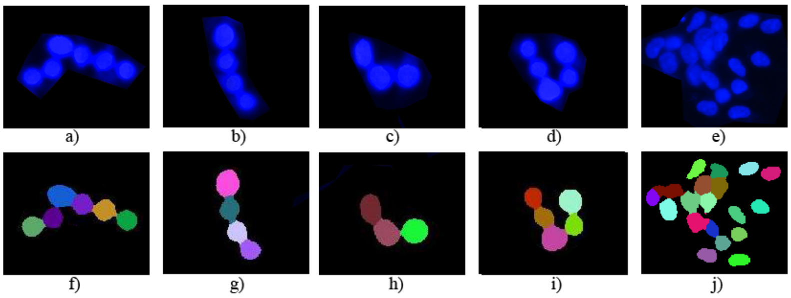

**Figure 7.** Several *declustering* examples for 6, 4, 3, 5 and 11 region clusters. a) through e) Original region patches. f) through j) Respective *declustered* patches.

Since not all biomedical researchers use the same annotation guidelines, two routines representing the two most popular techniques [16] were implemented:

- If the cell overlaps with one or more parasites;
- If a parasite is within a certain radius (~50% of the average cell radius).

Cellular and parasitic cytoplasmic membership was computed via set algebra on the region vectors. The radius was calculated using the Euclidean between the centers of mass and boundaries of each declustered region.

## 5 Results

To appraise the proposed method, the segmentation and classification steps were individually assessed. For this, two stress datasets (DSA, B) were built. Dataset A tested the segmentation step by presenting dim and out-of-focus conditions, while dataset B included highly clustering images, thus straining the classification step. Three biomedical researchers provided the ground truth by manually annotating both datasets.

Researchers were asked to carefully perform the annotations in separate days and double-check them, so as to minimize human error.

### 5.1 Segmentation

Since the method assumes that the cell/parasite identification process is not completely performed in the segmentation step, multi-nucleic regions were considered as well segmented results. A region was considered ill segmented if: a) it was not detected or b) its geometry was not correctly identified. Table 1 presents the obtained detection accuracies on both datasets, when compared with the ground truth.

Table 1. SEGMENTATION ACCURACY RATINGS

| | ***Macrophages*** | ***Parasites*** |
|---|---|---|
| Segmentation total ($DS_A$) | 3916 | 5257 |
| Ground truth ($DS_A$) | 4025 | 5572 |
| Accuracy percentage ($DS_A$) | 97.29 | 94.35 |
| Segmentation total ($DS_B$) | 4813 | 1832 |
| Ground truth ($DS_B$) | 5034 | 1981 |
| Accuracy percentage ($DS_B$) | 95.60 | 92.50 |

### 5.2 Classification Results

As dataset B possesses almost no testing data for either classifier due to its low clustering index, both classifiers were tested with data extracted from randomly selected regions in each of dataset A's images. Care was taken so that the data was divided, as equally as possible, between each class. The rule-based classifier's results are detailed bellow, in Table 2. No distinction was made between cellular or parasitic regions.

Table 2. RULE-BASED CLASSIFIER ACCURACY RATINGS

| **Class** | ***0 Error Margin (correct)*** | ***±1 Error Margin*** | ***±2 Error Margin*** | ***±3 Error Margin*** |
|---|---|---|---|---|
| 0 (noise) | 0.94 | 0.06 | 0.00 | 0.00 |
| 1 | 0.85 | 0.12 | 0.03 | 0.00 |
| 2 | 0.84 | 0.13 | 0.02 | 0.01 |
| 3 | 0.83 | 0.11 | 0.04 | 0.02 |
| 4 | 0.86 | 0.09 | 0.03 | 0.02 |
| 5 | 0.93 | 0.06 | 0.01 | 0.00 |
| 6 | 0.93 | 0.05 | 0.01 | 0.01 |
| 7 | 0.94 | 0.04 | 0.02 | 0.00 |
| 8 | 0.95 | 0.03 | 0.01 | 0.01 |
| 9 | 0.96 | 0.03 | 0.01 | 0.00 |

It is clear that as the number of nuclei increases, the classifier's error margins quickly decrease. The reason behind this is that, as the number of nuclei present in a region increases, the more the area value (normalized to the number of nuclei) approximates the distribution's mean value.

As previously mentioned, the log-likelihood ratios and their first and second order derivates were computed from mixture modeling 150 random clustered regions. These were then used to build several classifiers using Weka [15]. From these, Table 3 highlights the ones with the highest accuracy. All methods were validated through 2 and 10-fold cross validation (CV), as well as a 66% sequential split (SS). Ultimately, the SVM classifier was chosen due to its superior accuracy results. As with the RBC, care was taken so that the data was divided uniformly between classes.

Table 3. MACHINE-LEARNING CLASSIFIER ACCURACY RATINGS

| Validation Type | *C4.5 [19]* | *Best-First Search [19]* | *FFNN [19]* | *SVM [19]* |
|---|---|---|---|---|
| Sequential Split | 79.4% | 82.4% | 82.4% | **85.3%** |
| 10-fold Cross Validation | 79.4% | 67.6% | 70.6% | 76.5% |
| 2-fold Cross Validation | 69.2% | 69.2% | 61.5% | 69.2% |

### 5.3 Execution Times

In order to assess the method's real-time capabilities, the running times for each step were computed in both datasets. Table 4 compares the mean running times of the proposed method, measured in seconds, with the mean annotation time for each dataset across subjects. The mean number of regions for dataset A was ~203.5 and ~607.3 for dataset B. The obtained results show that, even with a computationally taxing approach, the method is capable of analyzing images considerably faster than human experts. Thus, real-time annotation proves a feasible reality, albeit with room for small code improvements, such as parallelization techniques.

Table 4. MEAN RUNTIMES AND STANDARD DEVIATION FOR THE PROPOSED METHODS VARIOUS STAGES (IN SECONDS). THE TESTS WERE RUN ON A LAPTOP WITH AN INTEL DUAL CORE T2410 2,0 GHZ CPU, 4 GB RAM, RUNNING WINDOWS 7 32-BITS. NO GRAPHICAL ACCELERATION WAS USED.

| | *Segmentation* | *Classification* | *Declustering* | *Total* | *Annotation* |
|---|---|---|---|---|---|
| Dataset A | 0.205 s | 1.156 s | 0.197 s | 276.3 s | 436.4 s |
| Dataset B | 0.287 s | 1.272 s | 0.316 s | 923.7 s | 1395.7 s |

## 6 Conclusions

This work has suggested a robust, automatic analysis methodology for cell and parasite detection in fluorescence microscopy imaging. The proposed method has shown itself robust to poor lighting conditions and high cluttering indexes, falling well within the error margins of expert biomedical researcher annotations.

The obtained results demonstrate this method is capable of performing the image analysis task adequately and in less time than a human expert. Being a computer program, the method also boasts from being immune to traditional human errors related to distraction, fatigue or subjectivity. Since these human errors are the major source of ambiguity in the manual annotation process, we consider our alternative to be a more suitable choice. This claim is supported with the fact that the method has a fixed error margin, meaning, the error does not randomly vary through time, as human errors do. Thus two or more drug trials can be safely compared as to their effectiveness, whereas if considering human error, the comparison requires validation through multiple confirmations.

Future work should focus on the methods used in the segmentation step, possibly employing mean shift or normalized cuts techniques, as well as increasing the training datasets of both classifiers. The built processing pipeline was made to be modular and applicable to other image types, thereby easily expandable to solve similar problems. This work has been successfully integrated with a pre-existing image annotation framework and is currently being used in the INEB/IBMC laboratories in Portugal.

## 7 References

[1] Ryan K. J., Ray C. G. 2004. Sherris Medical Microbiology. McGraw Hill. pp. 749–754.

[2] Myler P., Fasel N. 2008. Leishmania: After The Genome. Caister Academic Press.

[3] Jeronimo S. M. B., DeQueiroz-Sousa A., Pearson R.D. 2007. Leishmaniasis. In: Goldman L, Ausiello Deds. Cecil Medicine. 23rd Ed. Philadelphia, Pa: Saunders Elsevier:Ch: 369.

[4] Lichtman J. W. and Conchello J. A. 2005. Fluorescence Microscopy. Nature Publishing Group.

[5] Spring K. R. 2010. MicroscopyU: Introduction to Fluorescence Microscopy.

[6] Liao Q., Deng Y. 2002. An Accurate Segmentation Method For White Blood Cell Images. Proceedings IEEE International Symposium on Biomedical Imaging, pp. 245-248.

[7] Ficarra E., Cataldo S. D., Acquaviva A., Macii E. 2011. Automated Segmentation of Cells With IHC Membrane Staining. IEEE Transactions on Biomedical Engineering, Vol: 58, Issue: 5.

[8] Jiang K., Liao Q., Dai S. 2003. A Novel White Blood Cell Segmentation Scheme Using Scale-Space Filtering And Watershed Clustering. Proceedings of ICMLC.

[9] Park J. and Keller J. M. 1997. Fuzzy Patch Label Relaxation in Bone Marrow Cell Segmentation. International Conference on Computational Cybernetics and Simulation, pp: 1133–1138.

[10] Begelman G., Gur E., Rivlin E., Rudzsky M., Zalevsky Z. 2004. Cell Nuclei Segmentation Using Fuzzy Logic Engine. Proceedings IEEE International Conference on Image Processing.

[11] Yu W., Lee H. K., Hariharan S., Bu W., Ahmed S. 2008. Level Set Segmentation of Cellular Images Based on Topological Dependence. ISAVC.

[12] Yan P., Zhou X., Shah M., and Wong S. T. C. 2008. Automatic Segmentation of High-Throughput RNAi Fluorescent Cellular Images. IEEE Transaction On Information Technology In Biomedicine, Vol. 12, No. 1.

[13] Nobuyuki Otsu (1979). "A threshold selection method from gray-level histograms". IEEE Trans. Sys., Man., Cyber. 9 (1): 62–66.

[14] Freeman H. 1961. On the encoding of arbitrary geometric configurations. IRE Transactions on Electronic Computers, pp:260-268.

[15] Neal R. A., Croft S. L. 1984. An in-vitro system for determining the activity of compounds against the intracellular amastigote form of Leishmania donovani. Journal of Antimicrobial Chemotherapy, Vol. 14, Issue: 5, pp: 463-75.

[16] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

[17] Douglas Reynolds. Gaussian Mixture Models. MIT Lincoln Laboratory, MA 02140, USA.

[18] Gonzales and Woods. Digital Image Processing, 3rd Ed. (DIP/3e). 2008.

[19] Bishop C. M. 2007. Pattern Recognition and Machine Learning. Springer. ISBN: 0387310738.