Unsupervised segmentation of cervical cell images using Gaussian Mixture Model

Srikanth Ragothaman[1], Sridharakumar Narasimhan[1], Madivala G Basavaraj[1], Rajan Dewar[2]
[1]Indian Institute of Technology Madras, Chennai, India
[2]Department of Pathology, University of Michigan, Ann Arbor, US
srikanth.ragothaman@gmail.com,{sridharkrn,basa}@iitm.ac.in
rdewar@med.umich.edu

Abstract

Cervical cancer is one of the leading causes of cancer death in women. Screening at early stages using the popular Pap smear test has been demonstrated to reduce fatalities significantly. Cost effective, automated screening methods can significantly improve the adoption of these tests worldwide. Automated screening involves image analysis of cervical cells. Gaussian Mixture Models (GMM) are widely used in image processing for segmentation which is a crucial step in image analysis. In our proposed method, GMM is implemented to segment cell regions to identify cellular features such as nucleus, cytoplasm while addressing shortcomings of existing methods. This method is combined with shape based identification of nucleus to increase the accuracy of nucleus segmentation. This enables the algorithm to accurately trace the cells and nucleus contours from the pap smear images that contain cell clusters. The method also accounts for inconsistent staining, if any. The results that are presented shows that our proposed method performs well even in challenging conditions.

1. Introduction

Cervical cancer is one of the leading causes of cancer death in women. In several cases, symptoms usually show up in advanced stages of cancer where the treatment is non responsive. Screening at early stages is important in reducing fatalities due to cervical cancer. Pap smear test is the most popular screening technique to diagnose cervical cancer where the cervical cells are smeared onto a glass slide and observed under microscope to look for abnormality in nucleus and cytoplasm. The abnormal cells that are potentially precancerous are called dysplastic cells. The dysplastic cells appear to possess bigger and darker nuclei. Cytoscreeners look for dysplastic cells to decide whether further tests are required for diagnosis and treatment.

Manual screening of pap smear samples for cancer diagnostics is tedious and prone to errors. Moreover, the penetration of conventional cytology in low and middle income countries is low due to cost and manpower constraints. Therefore, there is enormous interest to develop suitable automatic screening systems that reduce human effort and increase adoption of cytology based techniques. In order to automate the screening process, nucleus and cytoplasm has to be segmented from pap smear image acquired via microscope. Hence accurate, unsupervised segmentation and classification of nucleus is critical to this automated diagnostic approach.

Nucleus and cytoplasm segmentation for single cell images have already been carried out earlier. Various methods have been developed to segment single cell images [5, 7, 10, 15, 16]. But generally cells appear in the form of clusters which possess local intensity variance within each cell makes the problem challenging. The segmentation of nucleus alone without its cytoplasm has been executed in the past using watershed segmentation, edge based contour fitting and active contour methods [2, 3, 12] which require higher contrast and intensity gradient difference around nucleus. Watershed based methods are highly subjected to preprocessing and requires user to select markers. Hierarchical clustering based segmentation of nucleus and cytoplasm have also been put forward to segment nucleus and cytoplasm in cells in the cases where several cells are present close to each other as clusters [8]. Most of methods mentioned above fails to perform well under poor staining conditions and overlapping of cells. Tareef *et al.,* [14] discuss methods to exactly segment nucleus and cytoplasm for overlapping cells. However, the applicability of the method to cell clusters is limited. Zhang *et al.,* [17] discussed cell segmentation and classification for particular staining technique but the method requires several parameters that need to be tuned. Some preliminary results with GMM have been

India Patent pending 201641012610

This work was partially supported by IIT Madras under the Socially Relevant Project Scheme

reported earlier [9]. Moreover segmenting cytoplasm and nucleus for each cell in a cluster of cells is difficult even for human eye as most of the edge information is hidden.

We propose a generic image processing method to segment nucleus and cytoplasm wherein the of pap smear images contain cells that overlap and are prepared under poor staining conditions [13]. Attempt to segment cytoplasm and demarcating for each cell is not made due to its complexities mentioned above. GMM is implemented to segment the nucleus and cytoplasm. In this method, pap smear images are assumed to be generated from mixture of Gaussians and each pixel is assigned to a class based on its weight associated with a component in mixture of distributions. The parameters for Gaussian distribution are calculated using expectation maximization (E-M) algorithm with the number of components determined heuristically. To eliminate the false negatives, this method is coupled with shape fitting of nucleus contours using Hough transform. Moreover, this method does not demand any preprocessing of images and can also be applied to cells that are clustered.

2. Dataset

Anonymized glass slides prepared using liquid based cytology (LBC) were provided for image analysis. The images are acquired using DinoLite digital microscope (camera sensor 1.3MP) of magnification ∼700x with image size 1280 x 1020. The dark spots in the image constitutes the nucleus and surrounding coloured area constitutes the cytoplasmic material of a cell. Some of the images were poorly or inconsistently stained (*Fig.1*) and had significant number of overlapping cells *Fig.2*.

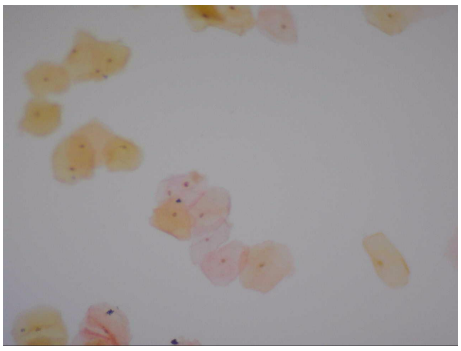


Figure 1. Clusters of cells with inconsistent staining

3. Methodology

3.1. Segmentation

Segmentation of nucleus and cytoplasm is the most important step for extracting cellular features from images and we propose to address this using a multi-step algorithm.

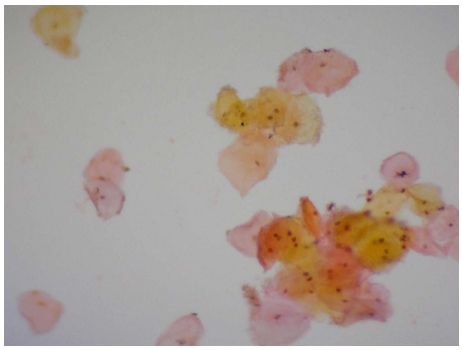


Figure 2. Image showing presence of overlapping cells

In the first step, GMM is implemented to segment the image roughly into certain number of classes. The number of classes is obtained heuristically using Akaike information criterion (AIC). After segmenting the image, further clustering of image regions is performed based on user defined parameters to reduce the final number of classes to three ie., nucleus, cytoplasm and background. This step is followed by morphological operations that removes over segmented nucleus based on its characteristics. Finally, the shape based identification of nucleus based on edge information is implemented to reduce false negative rates using Hough transform methods.

3.1.1 Gaussian Mixture Model

Gaussian Mixture Model (GMM) has been widely used for image segmentation [11]. Each pixel in the image is modelled as belonging to a class out of mixture of Gaussians. GMM algorithm estimates the probability of a pixel belonging to a class by modelling probability density functions (PDF) of the pixel's intensity values. The parameters of the PDF i.e., mean and covariance are evaluated using E-M algorithm.

3.1.2 Expectation Maximization (EM)

The EM algorithm is the common method of estimating parameters of GMM in which the maximum likelihood estimates are iteratively determined. The usual EM algorithm constitutes of E step and M step. E step computes the logarithmic likelihood of entire data set with K samples. M-Step finds the parameters by maximising the logarithmic likelihood function. The procedure is iterated till convergence. The mean and covariance matrix arrived at the coverging step is considered as the final model parameters. The image is quantized into m levels with mean calculated as intensity values.

3.1.3 Model Selection

One of the key concerns in probabilistic modelling is appropriate model selection, which in this case refers to the number of components in the mixture model. The number of Gaussian components to be estimated can be heuristically found using various information criteria techniques. Here AIC [1] is used to estimate the approximate number of components present in the data.

$$AIC = 2p - 2\ln(L)$$

where, p is the number of parameters and L is the maximum of likelihood value of model. AIC values are calculated for GMM ranging from 1 to 15 components and the model possessing minimum AIC value is chosen. From the *Fig.3*, it is clear that the AIC value is approximately the same for GMM with 8 upto 15 components. Hence to reduce the computational load and over fitting, it is appropriate to choose lesser number of components in the model. Moreover, the change in AIC value is less than 1% between the model with 15 components and model with 8 components.

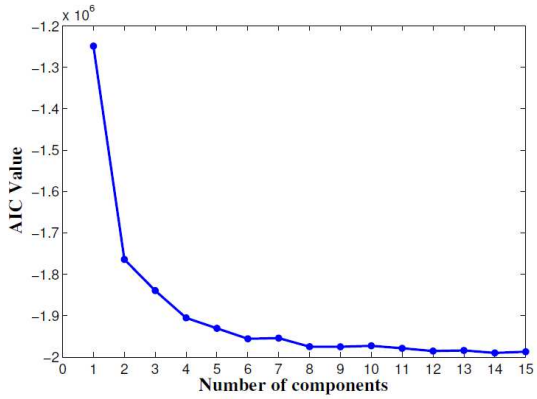


Figure 3. AIC vs Number of Components

3.1.4 Region Merging

The number of classes to be finally resolved is known *a priori* i.e., nucleus, cytoplasm and background. Since the regions in the image have higher variations within each class, applying the AIC criteria results in identifying a large number of classes. Hence appropriate merging of regions is essential to reduce the classes to 3. The classes are merged based on user defined criteria which can be tuned according to nature of image acquired. In case of poorly stained samples, these parameters can serve as a tool to avoid poor segmentation. The obtained image after clustering is converted to gray scale image. The intensities of quantized image are sorted in descending order. It is obvious that intensity of nucleus region is least and intensity of background is highest.

Consider a vector L of size m containing the quantized levels in descending order. L_{max} and L_{min} are the maximum and minimum level intensities. If

$$L(m) - L(m-1) < T_N$$

where T_N is the threshold for nucleus to be merged, then regions corresponding to last 2 levels are merged. If $L_{max} > T_B$ then, the region corresponds to L_{max} becomes single brighter background region where T_B is the minimum intensity threshold for the regions in the image to be labelled as background. If

$$L(m-(m-1)) - L(m-(m-2)) < T_{BD}$$

then $L(m-(m-1)) = 1$. If the difference between maximum and second maximum level is less than threshold T_{BD}, the region corresponding to those level are merged. The remaining regions are merged to become single region that results in segmenting cytoplasm. If $L_{max} - L_{min} < T_L$ then all regions are merged to single background layer ie., when the contrast is less than T_L, regions of cells are said to be absent and corresponding pixels are labelled as background. T_N, T_{BD} and T_L are usually fixed at $3-4\%$ of maximum intensity value in the image. T_B is approximately half the value of maximum intensity of image. These methods of merging with user defined parameters are on par with clustering techniques.

3.1.5 Identifying over segmented Nucleus

False segmentation of regions in image is likely to occur by implementing the above GMM-EM framework on pap smear images when intensity ranges of cytoplasm and nucleus overlap. This may occur due to poor imaging conditions (low light and issue with focus) and inconsistent staining of slides. Apriori knowledge of characteristics of nucleus are exploited to reduce the high false positive identification of nucleus. The nucleus shape is most likely circular but also occur in form of ellipses. Hence additional features such as area, eccentricity and major axis to minor axis ratio plays role in identifying the nucleus correctly and reducing the possibility of falsely segmented nuclei. Abnormally sized nuclei can be distinguished from falsely segmented nuclei by comparing the ratio of nucleus area to cytoplasm area. If the nucleus area, eccentricity and ratio of major to minor axis is within user specified range, then the nucleus is said to belong to set S. The darker regions that do not satisfy these conditions are eliminated. This results in higher accuracy in estimating the cytoplasmic area. The set S consists of all nucleus satisfying the conditions below.

$$N \in S \text{ if } C_1 \cap C_2 \cap C_3$$

$C1 : R_{min} < R < R_{max}$
$C2 : E_{min} < E < E_{max}$
$C3 : A_{min} < A < A_{max}$
Where, A, R, E are the area, major to minor axis ratio and eccentricity of nucleus. R_{min} and R_{max} denotes minimum and maximum of major to minor axis ratio, E_{min} and E_{max} denotes minimum and maximum of eccentricity and A_{min} and A_{max} denotes minimum and maximum of area of nucleus. R_{max} is typically between 4 to 6. In order to remove highly irregular shape particle E_{max} is usually chosen between 0.8 and 0.9. A_{max} is fixed at 5% of total cell size known apriori which depends on the magnification level of microscope. List of tuning parameters are listed in *Table. 1*

3.1.6 Shape based identification of nucleus

Debris present on a slide either due to environmental exposure or during sampling may account for darker pixels in the image as in *Fig.7* with similar size and shape of nucleus. This debris when darker than the true nucleus does have a negative effect on segmentation. Presence of these particles in the images forces the algorithm to falsely segment the nucleus as cytoplasm. Though tuning of T_N would solve this issue, manual tuning of each and every image acquired is not recommended due to random occurrence and unknown intensity information of debris. In such cases, shape information of nucleus is utilised to segment the nucleus that are not identified. The intensity gradient around the nucleus is relatively higher than the gradient between cytoplasm and background. The nucleus can be considered to be approximately circular object lying on the cytoplasmic background and Hough transform based methods [6] can be used to identify nucleus based on its shape. The image acquired is converted to gray scale image and edges are identified using canny edge detector [4]. This is followed by circular object detection using Hough transform. Since the edge detection is sensitive to noise, the sensitivity of Hough transform is kept low as to detect only sharp edges. The nucleus identified by Hough transform has to satisfy the user defined criteria mentioned above to reduce high false positive rates. Moreover the entire method is developed such a way that the false negative rate in identifying the nucleus is kept as low as possible.

Apart from above parameters, another information about nucleus is their spatial location with respect to cytoplasm. Nucleus is always present in the centre surrounded by cytoplasm. This prior information is exploited to avoid nucleus that are falsely segmented at edges and are present on the cytoplasmic borders.

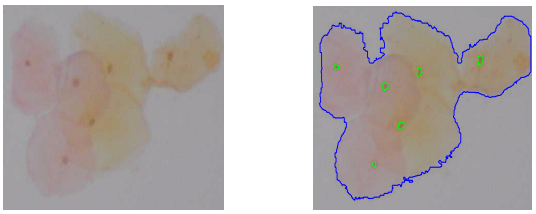
Table 1. List of Parameters

Parameters	Description
T_N	Difference between two lowest intensity levels
T_{BD}	Difference between two highest intensity levels
T_L	Difference between highest and lowest intensity levels
T_B	Minimum intensity to be labelled as background
R_{min},R_{max}	Minimum and maximum radius of nucleus
E_{min},E_{max}	Minimum and maximum eccentricity of nucleus
A_{min},A_{max}	Minimum and maximum area of nucleus

4. Results

The step by step procedure of algorithm with the results is illustrated. In the *Fig.5*, GMM segmentation of the the RGB image of overlapping cervical cells is done. Colour is quantized to m levels. Secondly, the RGB image converted to gray scale image to morphologically process the image to remove noise and oversized nucleus. The eccentricity and ratio of major axis to minor axis are also considered to remove over segmented areas. This followed by imposing a spatial constraint on nucleus position to eliminate nucleus present at vicinity of the cytoplasm. Dark regions indicate the nucleus, grey region indicates the cytoplasm area and white region denotes the background. In *Fig.6* illustration of cytoplasm and nucleus segmentation is made for whole field of view obtained. Contours of nucleus and cytoplasm indicate accurate segmentation. Implementation was carried out in MATLAB and associated toolboxes.

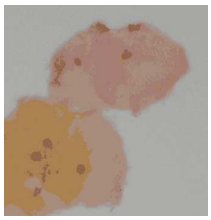
This method works well even for poorly stained slides as illustrated in *Fig.4*. Consistent staining is always not expected even in automated staining machines. At certain instances, the intensity distribution of the nuclear region may not account for all the nuclei present due to inconsistent staining and inherent errors propagated during image acquisition. With the help of shape based identification of nucleus, the nucleus detection accuracy is increased as illus-



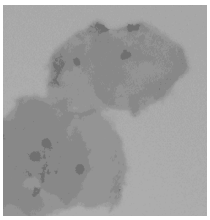
(a) Original Image (b) Image showing segmented regions

Figure 4. Segmentation of cell clusters even with poor staining

(a) Original image



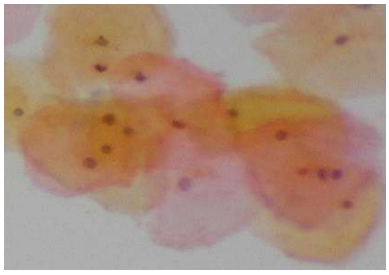
(b) Image quantized to m levels



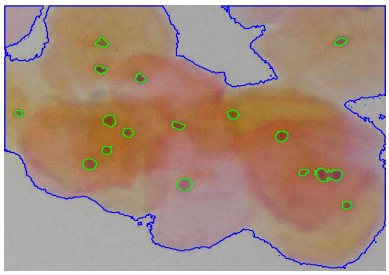
(c) Gray scale image of colour quantized image

(d) Segmented image with dark regions indicating nucleus and gray region indicating cytoplasm

Figure 5. Step by Step illustration of proposed algorithm



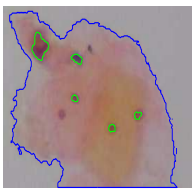
(a) Original image



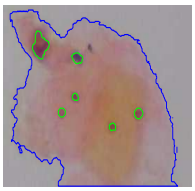
(b) Image showing segmented regions

Figure 6. Segmentation for clusters of cells

(a) Original Image



(b) Segmented image



(c) Identifies nucleus with shape

Figure 7. Effect of shape based identification of nucleus

trated in *F*ig.7. Though debris is misclassified as nucleus in *Fig.7*, overall methods reduces the false negative rate. The number of components chosen in model selection step does not have any effect on segmentation when the components chosen is greater than 8, though false segmentation arises when the number of components chosen is significantly low. The number of components is selected between 8 and 10 to account the trade off between over fitting and computational load.

5. Performance measures

Performance is measured by examining the algorithm results against ground truth. The algorithm is evaluated on the images containing cells with a total of 148 nuclei. Each of the images are manually segmented into nucleus and cytoplasm for evaluation purpose with coordinates of the center of the nucleus marked and considered as ground truth. Precision and recall are treated as performance measures for evaluating nucleus identification that are commonly used in field of information retrieval and also used in object based evaluation procedures.

$$\text{Recall} = \frac{\text{No.of correctly detected Nucleus}}{\text{No.of Nucleus present in image}}$$

$$\text{Precision} = \frac{\text{No.of correctly detected Nucleus}}{\text{Total number of detected nucleus}}$$

Recall was found to be 94.90 and precision to be 91.46. The under segmentation of remaining nucleus may be attributed to highly overlapping nature of cells where some

potential gradient information are lost underneath the overlapped cells. The accuracies of segmented contours of cytoplasm were also quantified using Dice similarity Coefficient (DSC). A_1 is the area segmented by proposed algorithm and A_2 is the ground truth area manually delineated. This measure is evaluated only for 20% of the image dataset acquired.

$$DSC = 2\frac{|A_1 \cap A_2|}{|A_1|+|A_2|}$$

Table 2. DSC measures

	$\mu_{DSC} \pm \sigma_{DSC}$	
	Nucleus	Cytoplasm
Consistent staining	0.86±0.03	0.96 ±0.03
Inconsistent staining	0.84±0.02	0.92±0.02

DSC values greater than 0.7 are considered satisfactory [18]. Given the limitations of the microscope and inconsistent staining conditions, the performance of the proposed method is acceptable.

6. Conclusion

The results shows the suitability of our proposed framework for automatic analysis of cervical cell images. The image is assumed to be generated from mixture of Gaussian and the parameters are found using E-M algorithm. Undersegmentation of nucleus is avoided using Hough transform based ellipse fitting methods to segment nucleus. We expect a very high accuracy for images when acquired through highly resolved optical microscope. This method even performs well for cells appearing in form of clusters and poorly stained slides.

References

[1] H. Akaike. *Selected Papers of Hirotugu Akaike*, chapter Information Theory and an Extension of the Maximum Likelihood Principle, pages 199–213. Springer New York, New York, NY, 1998.

[2] P. Bamford and B. Lovell. Unsupervised cell nucleus segmentation with active contours. *Signal Processing*, 71(2):203–213, Dec. 1998.

[3] C. Bergmeir, M. G. Silvente, and J. M. Bentez. Segmentation of cervical cell nuclei in high-resolution microscopic images: A new algorithm and a web-based software framework. *Computer Methods and Programs in Biomedicine*, 107(3):497 – 512, 2012.

[4] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, June 1986.

[5] Y. F. Chen, P. C. Huang, K. C. Lin, H. H. Lin, L. E. Wang, C. C. Cheng, T. P. Chen, Y. K. Chan, and J. Y. Chiang. Semi-automatic segmentation and classification of pap smear cells. *IEEE Journal of Biomedical and Health Informatics*, 18(1):94–108, Jan 2014.

[6] R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, Jan. 1972.

[7] J. Fan, R. Wang, S. Li, and C. Zhang. Automated cervical cell image segmentation using level set based active contour model. In *Control Automation Robotics Vision (ICARCV), 2012 12th International Conference on*, pages 877–882, Dec 2012.

[8] A. Gençtav, S. Aksoy, and S. Önder. Unsupervised segmentation and classification of cervical cell images. *Pattern Recognition*, 45(12):4151–4168, Dec. 2012.

[9] G. K. Lakshmi and K. Krishnaveni. Multiple feature extraction from cervical cytology images by gaussian mixture model. In *Computing and Communication Technologies (WCCCT), 2014 World Congress on*, pages 309–311, Feb 2014.

[10] K. Li, Z. Lu, W. Liu, and J. Yin. Cytoplasm and nucleus segmentation in cervical smear images using Radiating GVF Snake. *Pattern Recognition*, 45(4):1255–1264, Apr. 2012.

[11] H. Permuter, J. Francos, and I. Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4):695 – 706, 2006. Graph-based Representations.

[12] M. E. Plissiti, C. Nikou, and A. Charchanti. Watershed-based segmentation of cell nuclei boundaries in pap smear images. In *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on*, pages 1–4, Nov 2010.

[13] R. Srikanth, S. Narasimhan, B. Gurappa, and R. Dewar. Methods and apparatus for analyzing cytological specimens. India patent pending 201641012610, April 2016.

[14] A. Tareef, Y. Song, W. Cai, D. D. Feng, and M. Chen. Automated three-stage nucleus and cytoplasm segmentation of overlapping cells. In *Control Automation Robotics Vision (ICARCV), 2014 13th International Conference on*, pages 865–870, Dec 2014.

[15] M.-H. Tsai, Y.-K. Chan, Z.-Z. Lin, S.-F. Yang-Mao, and P.-C. Huang. Nucleus and cytoplast contour detector of cervical smear image. *Pattern Recognition Letters*, 29(9):1441–1453, July 2008.

[16] S. F. Yang-Mao, Y. K. Chan, and Y. P. Chu. Edge enhancement nucleus and cytoplast contour detector of cervical smear images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(2):353–366, April 2008.

[17] L. Zhang, H. Kong, C. Ting Chin, S. Liu, X. Fan, T. Wang, and S. Chen. Automation-assisted cervical cancer screening in manual liquid-based cytology with hematoxylin and eosin staining. *Cytometry Part A*, 85(3):214–230, 2014.

[18] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer. Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE Transactions on Medical Imaging*, 13(4):716–724, Dec 1994.