Quantitative Validation of Segmentation Methods of Cervical Cytology Images

G.Karthigai Lakshmi
Department of Computer Science
V.V.Vanniaperumal College for Women, Virudhunagar, Tamilnadu, India.

K.Krishnaveni
Department of Computer Science
Sri Ramasamy Naidu Memorial College, Sattur, Tamilnadu, India.

Abstract- Feature extraction is the crucial process in deciding the stage of cervical cancer. Cervical Cytology images obtained from Pap smear test are preprocessed and segmented to extract relevant features of nucleus and cytoplasm. The key focus of this paper is to validate the output of two segmentation methods - Color K-means clustering and Gaussian Mixture Model. Jaccard and Dice Similarity Coefficients are used to evaluate the quality of segmentation against the ground truth. Box plot is used to point up the minimum, maximum and mean values of the above mentioned coefficients for the seven classifications of the cervical cytology images taken from the Herlev University database. Based on the comparison, Gaussian Mixture model which works on gray scale images is found to yield better result than the color K-means clustering segmentation method which uses RGB color images.

Keywords – Cervical Cytology images, Color K-means Clustering, Gaussian Mixture Model, Similarity coefficients

I. Introduction

Cervical cancer is the most prevalent cancer affecting women. Human Papilloma Virus (HPV) infection in women causes cervical cancer. Pap smear test is a diagnosis method used for the prediction of cervical cancer. Cervical cancer is curable if diagnosed at an early stage by Pap smear test or Thin Prep test. The mortality rate due to cervical cancer is high in developing and underdeveloped countries due to ignorance of the symptoms and nature of this disease, India has a population of 432.20 million women aged 15 years and older who are at risk of developing cervical cancer.[1] Having regular Pap tests annually, above the age of 40, is the best way for women to protect themselves against cervical cancer. Pap test identifies the pre-malignant cells in cervix of the uterus, which may progress to malignant cells. Automated analysis of cervical cancer yields a more precise result compared to human diagnosis. Cervical cytology images obtained from the microscope are preprocessed by the computer and an image with enhanced quality in terms of signal-to-noise ratio and contrast is obtained. The preprocessed image is then segmented into background, cytoplasm and nucleus. The features of the nucleus and cytoplasm are employed to determine the severity and stage of cancer. The accurate diagnosis of the disease relies on the features extracted and this extraction in turn depends on segmentation techniques used. This paper compares two segmentation methods – K- means clustering and Gaussian Mixture model of segmentation. The pre-eminence of Gaussian Mixture model is proved by similarity coefficients. In this paper, Pap Smear images from database of Herlev University[2], Denmark have been used.

The rest of the paper is organized as follows. Proposed segmentation algorithms are explained in section II. Calculation of Similarity Coefficients is illustrated in section III. Brief description of the box plot is given in section IV. Experimental results are presented in section V. Concluding remarks are given in section VI.

II. Proposed Algorithms

A. K-means Clustering algorithm

Segmentation methods can be classified as region based, boundary based and hybrid methods. Clustering is a way to separate objects in images. K-means clustering is a region based clustering method. It is a simple, unsupervised method with less time complexity. K-means clustering is suitable for medical image segmentation, as the number of clusters (K) is known for most of the images of particular regions of human anatomy.

In cervical cytology images, the number of clusters is 3, representing the regions of background, cytoplasm and nucleus. Since RGB images obtained from Pap smear test are used, nucleus and cytoplasm may be divided into further clusters due to inhomogeneity in color of those regions. K-means algorithm classifies the input image into multiple regions based on their distance from each other. The distance metrics is the difference between the intensity value of a pixel and the cluster mean. The algorithm assumes that the image forms a vector space and tries to identify clusters in it, based on intensity comparison. The steps involved in the color K-means clustering are depicted in Figure 1.

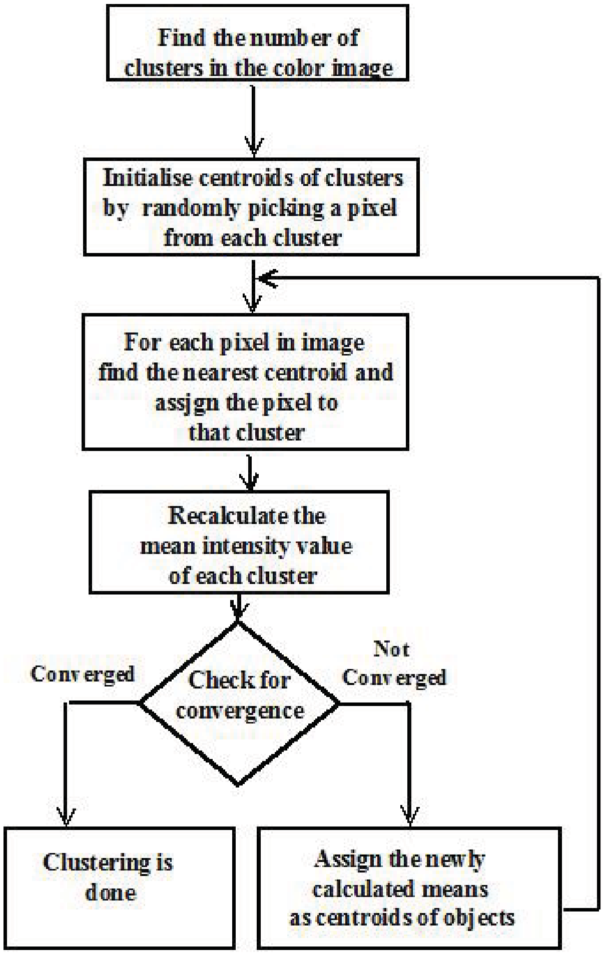

Find the number of clusters in the color image
Initialise centroids of clusters by randomly picking a pixel from each cluster
For each pixel in image find the nearest centroid and assjgn the pixel to that cluster
Recalculate the mean intensity value of each cluster
Check for convergence
Converged
Not Converged
Clustering is done
Assign the newly calculated means as centroids of objects

Figure 1. Block diagram of Color K-means Clustering algorithm

K-means algorithm needs an initial value of number of clusters in the color cervical cytology image. Using statistical packages, the number of clusters is calculated. Based on the number of clusters, initially pixels are selected at random and their intensity values are assigned are initial means of clusters. The pixels in the image are then grouped into clusters by checking the proximity of their intensity values to the means of clusters. For the newly formed clusters, mean values are calculated. If the differences between the previous and current values are minimal, then clustering is completed. If the differences are high, clustering is done again.

B. Gaussian Mixture Model algorithm

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous features in biomedical images. The cervical cytology image is an RGB image with three different regions viz., background, cytoplasm and nucleus. Such images contain pixels with varying intensity values for the three regions. Therefore, instead of a single Gaussian assumption for the probability function of the pixel intensity, a mixture of three Gaussian distributions was used to model the pixel value. The three distributions correspond to the three regions in the image. Mixture of Gaussians is extremely popular in image segmentation and background separation. A block diagram of the Gaussian Mixture model is given in Figure 2.

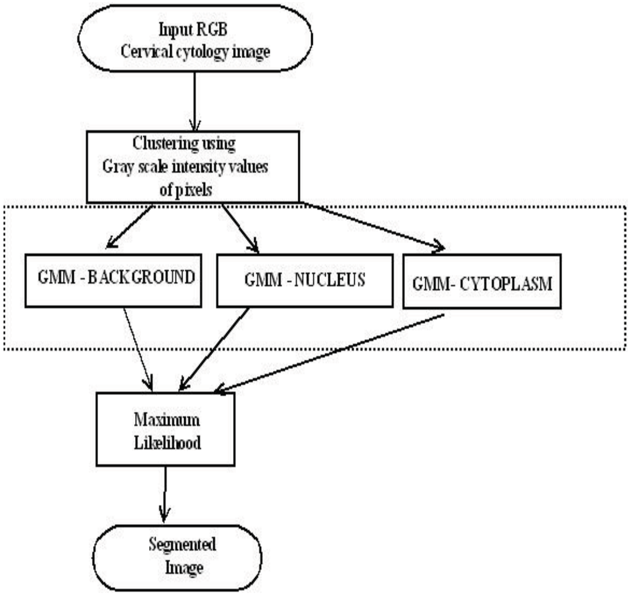


Figure 2. Block diagram of Gaussian Mixture Model algorithm

III. Calculation of Similarity Coefficients

Abundant similarity indices have been proposed to measure the degree to which the segmented image matches with the ground or gold truth. In this paper, Jaccard and Dice co-efficient have been used to measure the efficiency of the segmentation algorithms.

If 'A' is the ground truth image and 'B' is the segmented image obtained by the above segmentation algorithms, then the Jaccard similarity coefficient is defined as the the intersection of the images divided by the union of the images A and B.

$$Jaccard\ Index = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

The Dice similarity coefficient is defined as twice the intersection of the images divided by the summation of the images A and B. The intensity values of the gray scale images are used in the calculations.

$$Dice's\ coefficient = \frac{2|A \cap B|}{|A| + |B|} \qquad (2)$$

IV. DISPLAY OF DISTRIBUTION OF DATA USING BOXPLOT

The box plot [3] is a standardized way of displaying distribution of data based on the five summary statistics criteria: minimum, first quartile, median, third quartile, and maximum. In box plot, the central rectangle spans the the *interquartile range* (first quartile to the third quartile) as shown in figure 3. A line segment inside the rectangle shows the median and "whiskers" above and below the box show the locations of the minimum and maximum. Outliers are either 3×*IQR* or more above the third quartile or 3×*IQR* or more below the first quartile and are shown using a + symbol. An example box plot is given in Figure 3.

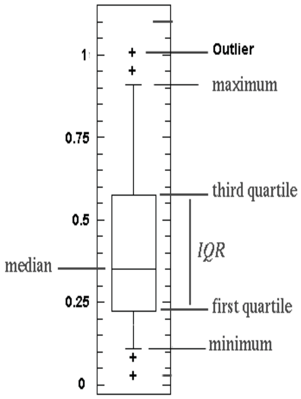


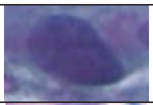
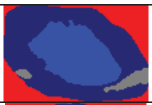
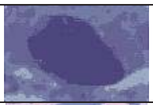
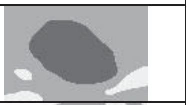
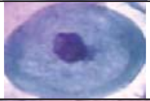
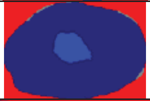
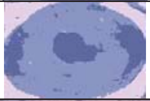
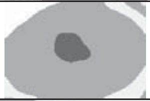
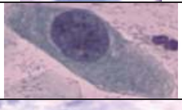
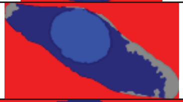
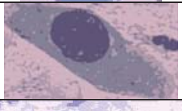
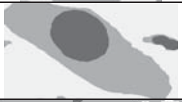
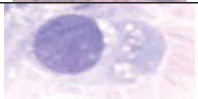
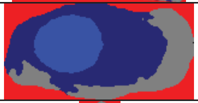
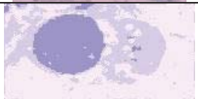
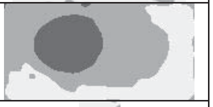
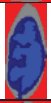
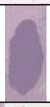
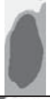
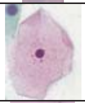
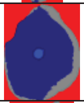
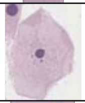
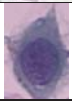
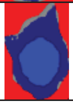
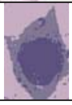
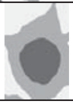
Figure 3. Components of simple Box plot

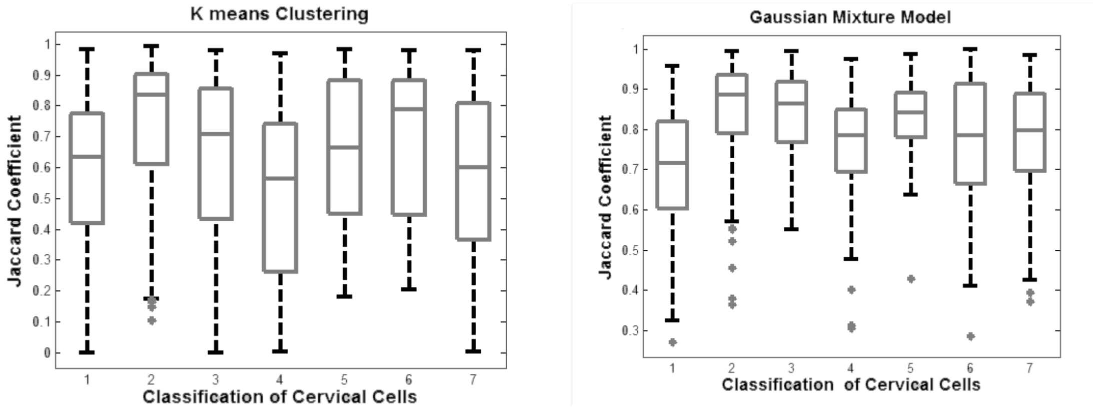
The values obtained for Jaccard and Dice similarity coefficients for the segmentation methods Color K-means clustering and Gaussian Mixture model are plotted as Box plots for ease of comparison.

V. EXPERIMENT AND RESULT

The test set for this evaluation experiment is taken from the Herlev University database. The database contains seven classes of images as normal superficial, normal intermediate, normal columnar, light dysplastic, moderate dysplastic, severe dysplastic and carcinoma in situ. The first three classes are normal cells and the remaining four classes are malignant cells. 917 images of seven classifications were segmented using K-means clustering and Gaussian Mixture Model. The gold truth is the results obtained using manual segmentation.

Table - 1 Results of Segmentation methods

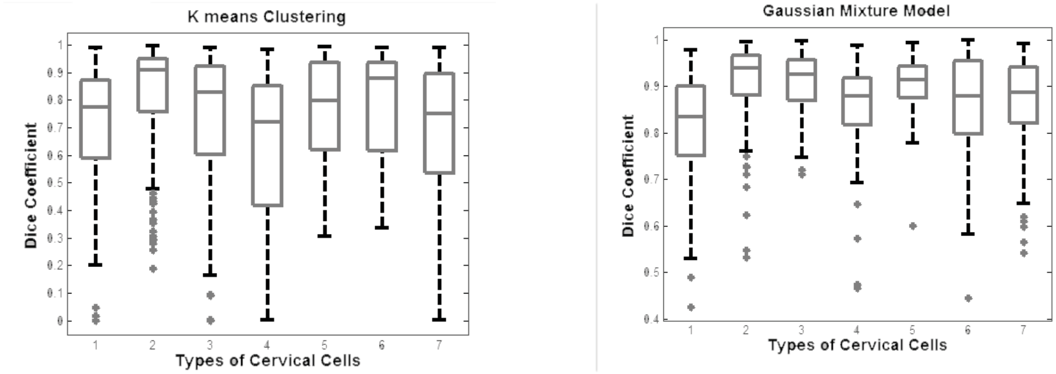
Classification of cells	Original image	Manually segmented Ground truth image	Image obtained from color K-means clustering	Image obtained from Gaussian Mixture model
Severe dysplastic				
Normal intermediate				
Light dysplastic				
Normal columnar				
Carcinoma in situ				
Normal superficial				
Moderate dysplastic				



(a) K-Means clustering (b) Gaussian Mixture model

1. Normal Superficial 2.Normal intermediate 3. Normal Columnar 4. Light dysplastic
5. Moderate dysplastic 6. Severe dysplastic 7. Carcinoma in situ

Figure 4. Box plot of Jaccard coefficient obtained for seven classes of images



(a) K-Means clustering (b) Gaussian Mixture model

1. Normal Superficial 2.Normal intermediate 3. Normal Columnar 4. Light dysplastic

5. Moderate dysplastic 6. Severe dysplastic 7. Carcinoma in situ

Figure 4. Box plot of Dice coefficient obtained for seven classes of images

Table - 2 Comparison of results - Jaccard coefficient

Jaccard coefficient				
Classification of Images	**K-means Clustering**		**Gaussian Mixture model**	
	Maximum	**Mean**	**Maximum**	**Mean**
Normal superficial	0.981838	0.588345	0.957895	0.698245
Normal intermediate	0.991774	0.734805	0.993208	0.849562
Normal columnar	0.980041	0.646589	0.994685	0.841540
Light dysplastic	0.96852	0.526832	0.987784	0.825611
Moderate dysplastic	0.983855	0.635488	0.987784	0.825611
Severe dysplastic	0.979778	0.659965	0.998316	0.781306
Carcinoma in situ	0.980437	0.587553	0.983244	0.778937

Table - 2 Comparison of results – Dice coefficient

Dice coefficient				
Classification of Images	**K-means Clustering**		**Gaussian Mixture model**	
	Maximum	**Mean**	**Maximum**	**Mean**
Normal superficial	0.990836	0.719682	0.978495	0.813966
Normal intermediate	0.995871	0.813411	0.996593	0.913205
Normal columnar	0.989922	0.750195	0.997335	0.910136
Light dysplastic	0.984009	0.652465	0.993855	0.853404
Moderate dysplastic	0.991862	0.749389	0.993855	0.901165
Severe dysplastic	0.989786	0.729092	0.999157	0.868862
Carcinoma in situ	0.990122	0.708157	0.991551	0.868858

Table 1 shows the results obtained for the Jaccard coefficient and Table 2 shows the results obtained for Dice coefficient. The comparison of the results shows that Gaussian mixture model yields better result than K-means clustering. For the normal superficial cell, the ratio of nucleus area to cytoplasm area is very small. The size of noisy bodies in these cells is greater than the size of nucleus in some cases. So the mean value of the Jaccard and Dice coefficients deviate from other classes to certain extent.

VI .CONCLUSION

This paper dealt with two segmentation method K-means Clustering and Gaussian mixture model for segmenting cervical cytology imagers. Both methods effectively segmented the three components of the image as background, cytoplasm and nucleus of the seven classes of cervical cytology images taken from the Herlev hospital database. Based on the comparison between the two methods by Jaccard and Dice similarity coefficients Gaussian Mixture model produces better results than K-means Clustering method. Similarity between hold truth images given by manual segmentation and Gaussian Mixture model is upto 99% in all the classes.

REFERENCES

[1] http://www.hpvcentre.net/statistics/reports/IND.pdf.

[2] http://labs.fme.aegean.gr/decision/downloads.

[3] http://www.physics.csbsju.edu/stats/box2.html

[4] G.K.Lakshmi,K.Krishnaveni",Multiple feature extraction from Cervical cytology images by Gaussian mixture model " , in *Proceedings of the World Congress on Computing and Communication Technologies(WCCCT) , IEEE Conference Publications(2014) pp.309-311*

[5] Mark Polak, Hong Zhang , Minghong Pi,"An evaluation metric for image segmentation of multiple objects", *Image and Vision Computing 27 (2009) 1223–1227*

[6] D. Martin, C. Fowlkes, D. Tal, J. Malik, "A database of human segmented natural images and its application to evaluating algorithms and measuring ecological statistic"*, ICCV (2001) 416–423*

[7] Jaime S. Cardoso, Luis Corte-Real, "Toward a generic evaluation of image segmentation", *IEEE Transactions on Image Processing 14 (11) (2005) 1773– 1782.*