SEGMENTATION OF LUNG GLANDULAR CELLS USING MULTIPLE COLOR SPACES

Sajith Kecheril S[1], D Venkataraman[1], J Suganthi[2] and K Sujathan[3]

[1]Department of Computer Science and Engineering, Amrita Viswa Vidyapeetham, Coimbatore, India
isajith@gmail.com, d_venkat@cb.amrita.edu.
[2] Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology, Coimbatore, India
sugi_jeyan@yahoo.co.in
[3] Department of Pathology, Regional Cancer Center, Thiruvananthapuram, India
ksujathan@gmail.com

ABSTRACT

Early detection of lung cancer is a challenging problem, the world faces today. Prior to classify glandular cells as malignant or benign a reliable segmentation technique is required. In this paper we present a novel lung glandular cell segmentation technique. The technique uses a combination of multiple color spaces and various clustering algorithms to automatically find the best possible segmentation result. Unsupervised clustering methods of K-means and Fuzzy C-means were used on multiple color spaces such as HSV, LAB, LUV, xyY. Experimental results of segmentation using various color spaces are provided to show the performance of the proposed system.

KEYWORDS

K-means, Fuzzy C-means, Color space, Sputum cytology

1. INTRODUCTION

In today's world, one of the major threats the mankind face is the mortality due to cancer. Among all cancers, lung cancer is the most prominent type among males. There is a need for a system which can detect the presence of malignancy at early stages. With the available man power present today it is extremely difficult to screen the entire population. So there requires an automated system which can screen all eligible people. Unlike other modalities, microscopic image analysis poses challenges in different dimension. Right from the slide preparation care has to be taken for satisfactory results. The chemical composition of reagents used for preserving and staining greatly affect the color and morphology of the exfoliated cells. Differential staining of slides may result in non uniform color values making the algorithm to be intensity invariant. Malignant cells are characterized by the change in the nuclear appearance. This paper attempts to find a proper segmentation of glandular cell nuclei which is subsequently used to extract features and used for classification.

Segmentation of nuclear region is a vital step in most of the image analysis of cytology images. Here a novel approach towards the segmentation of glandular cells is presented. The pap stained

sputum samples are digitized using a digital microscope. The staining of the cellular materials varies widely in conventional pap staining. The staining color varies from slide to slide depending on various parameters like batch of stain prepared, age of cell, type of cell and also depends on the person preparing the slide.

Various authors have tried several methods for segmentation of nuclear regions from cytological images. Walker et al [1], Thiran et al [2] used morphological operations for segmentation, while Schnorrenberg et al [3], Wu et al [4] used iterative methods. Sammouda et al [5] used Hopefield Neural Network (HNN) on RGB and HSV color spaces to extract adenocarcinoma nuclei of lung, Tanaka et al [6],Wittke et al [7] used thresholding for segmentation on grayscale images, Meas-Yedid et al [8] used Markov Random Field (MRF) clustering on RGB color space, Begelman et al [9] used fuzzy logic on CIELAB color space, Hu et al [10] used a variant of active contour on grayscale image, Phukpattaranont et al [11] applied neural network on RGB image for segmentation. Ko et al [12] used adaptive window based on local contrast on grayscale image, Palacios et al [13] used discrete wavelet transform and K-means clustering on HSV, LUV and YIQ color spaces. Taher et al [14] used HNN and FCM to segment lung cells on RGB image, Plissiti et al [15] used grayscale morphological reconstruction and connected regional minima along with FCM and SVM to detect nuclei in pap smear images. Wang et al [16] applied graph cut followed by active contour for nuclear segmentation.

In RGB color space the color intensity values are highly correlated so in order to extract valuable information it has to be decorrelated. So various authors have used different color spaces for information retrieval Huang et al [17], Chen et al [18], Zhao et al [19],Yuzheng et al [20], Anari et al [21] used HSV color space for segmentation. Combination of color spaces [22-28] also yielded good results. Unsupervised learning of K-means clustering [29] and Fuzzy C-means [15] were used for cell segmentation.

2. MATERIALS AND METHODS

2.1. Materials

The sputum cytology images used in this work were provided by Pathology department, Regional Cancer Center Thiruvananthapuram, India. The nature of cells was confirmed by expert cytopathologist. The images were captured by using a trinocular microscope fitted with a digital camera. The images were captured at 40X magnification level. The images are of 24 bit color depth with a spacial resolution of 3264x2448 pixels.

2.2. Clustering Methods

In the proposed algorithm, two clustering techniques such as K-means and Fuzzy C-means are used for segmentation.

2.2.1. K-means Clustering

K-means is an unsupervised clustering technique in which m data points with n dimension is grouped into k separate groups, $S_i, i = 1, \cdots, k$ [30]

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

(1)

The grouping is done by minimizing the distance between data point and the centroid of the corresponding cluster to which it falls to. X_j the data point, μ_i the centroid of the ith cluster, $\|\cdot\|^2$ is the l2 norm of a vector. The algorithm is run multiple times so that no more change to the data point happens. For this paper, $k=2$ is chosen, one for foreground and the other for background.

2.2.2. Fuzzy C-means Clustering

The K-means clustering assigns distinctly one group to each data point. In many applications, especially medical application, it may not be clear to which group a particular point belongs. In fuzzy C-means a membership value is assigned to the data point for each cluster. A data point near the center of cluster has a high membership value to that particular cluster than a point near the edge of the cluster. The objective function is as

$$E = \sum_{j=1}^{c} \sum_{i=1}^{N} \mu_{ij}^{k} \|x_i - c_j\|^2 \tag{2}$$

μ_{ij} is the fuzzy membership of the data point x_i, c_j is the centroid, k is the constant defining fuzzyness of the cluster. After each iteration the membership function and cluster centroid is updated as

$$\mu_{ij} = \frac{1}{\sum_{m=1}^{C} \left(\frac{\|x_j - c_i\|}{\|x_j - c_m\|^{2/(k-1)}} \right)} \tag{3}$$

$$C_i = \frac{\sum_{j=1}^{N} \mu_{ij}^{k} x_j}{\sum_{j=1}^{N} \mu_{ij}^{k}} \tag{4}$$

2.3. Color Space Conversion

The RGB image contains pixel values which are highly correlated, so processing it directly results in redundant information processing. To uncorrelate this data we use other color spaces like HSV, LUV, LAB and xyY.

2.3.1. RGB to HSV

The input RGB image is normalized before converting to HSV color space image using the following method [31].

$$N = max(R, G, B) \tag{5}$$

$$n = min(R, G, B) \tag{6}$$

Let, $\quad r = \dfrac{N-R}{N-n}; g = \dfrac{N-G}{N-n}; b = \dfrac{N-B}{N-n};$ \tag{7}

Value is $\quad V = max(R, G, B)$ \tag{8}

Saturation is $\quad S = \dfrac{N-n}{n}$ \tag{9}

$$
\begin{aligned}
H &= \frac{\pi}{3}(b-g), \text{ if } R=N \\
&= \frac{\pi}{3}(2+r-g), \text{ if } G=N \\
&= \frac{\pi}{3}(4+g-r), \text{ if } B=N
\end{aligned}
$$

Hue is (10)

$$
H = \frac{H}{2\pi}
$$

Normalized hue is (11)

2.3.2. RGB to CIELAB

CIELAB color space is derived from CIEXYZ color space. In CIEXYZ all visible colors are represented using only positive values. The color conversion matrix is as,

$$
\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 3.240479 & 1.53715 & 0.498535 \\ -0.969256 & 1.875992 & 0.041556 \\ 0.055648 & -0.204043 & 1.057311 \end{bmatrix} \times \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}
$$

(12)

The CIELAB color components are given by

$$
A = 500 \left[h\left(\frac{X}{X_w}\right) \quad h\left(\frac{Y}{Y_w}\right) \right]
$$

(13)

$$
B = 200 \left[h\left(\frac{Y}{Y_w}\right) \quad h\left(\frac{Z}{Z_w}\right) \right]
$$

(14)

$$
\begin{aligned}
h(q) &= \sqrt[3]{q}, , \text{ if } q > 0.008856 \\
&= 7.787q - 16/116, \text{ if } q < 0.008856
\end{aligned}
$$

(15)

Where X_w, Y_w and Z_w are the white tristimulus values of a perfectly reflecting diffuser under CIE standard D65 illumination [32].

2.3.3. XYZ to xyY

CIExyY is derived from CIEXYZ color space. The conversion is as follows

$$
x = \frac{X}{X+Y+Z}
$$

(16)

$$
y = \frac{Y}{X+Y+Z}
$$

(17)

$$
Y = Y
$$

(18)

2.3.4. CIEXYZ to CIELUV

LUV color space is derived from XYZ color space. The conversion is as follows

$$
u' = \frac{4X}{X+15Y+3Z}
$$

(19)

150

$$v' = \frac{9Y}{X + 15Y + 3Z} \tag{20}$$

$$
\begin{aligned}
L^* \quad &- \quad 116 \left(\frac{Y}{Y_w}\right)^{\frac{1}{3}}, \text{ if } \frac{Y}{Y_w} > \left(\frac{6}{29}\right)^3 \\
&= \quad \left(\frac{29}{3}\right)^3 \left(\frac{Y}{Y_w}\right), \text{ if } \frac{Y}{Y_w} \leq \left(\frac{6}{29}\right)^3
\end{aligned}
\tag{21}
$$

$$u^* = 13L^*(u' - u'_w) \tag{22}$$

$$v^* = 13L^*(v' - v'_w) \tag{23}$$

3. CELL NUCLEI SEGMENTATION

The system architecture for the proposed method to classify the sputum cytology cells is shown in Figure 1. It includes pre-processing of Pap stained sputum cytology images followed by segmentation of cell nuclei.

Figure 1. Diagram of the proposed system

3.1. Input Image

The sputum samples taken by cytotechnologists are preserved so that the cellularity of the nuclei is not lost. Then slides are prepared and stained using Pap stain. This is a differential stain in which the nuclei and cytoplasm region gets stained in different color. Nuclei region being more with dense materials will absorb more stain and appear much darker in color. The stained slide is placed under a trinocular light microscope and images are taken. The digital camera produces a high resolution image of 8 mega pixel (3264 X 2448) size. The slides are magnified at 40X zoom level and images are taken. The images are stored on the workstation in RGB 24 bit color format. Images of both malignant and benign slides are taken. The classification of the slides as benign and malignant is done by experienced cytopathologists.

3.2. Pre Processing

The high resolution input image is very difficult to process as such so in this stage the image is down sized to 1/8th of the original size (408 X 306). Local minima are found on the smoothed version of the image using Blom's method as used by Kuijper [33]. Once the local minima are identified then the coordinate in the resized image is mapped on to the original image. This is done by multiplying the x and y coordinates values with 8. Thus on the actual image coordinate a 128 X 128 sized image region is cropped out and is stored. The image region may contain region without cellular material i.e., background region. This has to be eliminated for further analysis. For this the average intensity value of the cropped region is calculated, if the value falls above 200 it is assumed to be background since nuclear region being darkly stained will have a low

average intensity only. The sputum cytology images are infiltrated by patches of mucus which also absorbs the stain. One more condition is given to filter out such region. The standard deviation of the extracted region is calculated if it is above 30 then there is an adequate distribution of dark and bright patches are present which is typical of nuclear region. All the cropped patches are saved. The sputum sample contains a lot of artifacts like inflammatory cells, leukocytes etc which is characteristic of the sputum sampling. These artifact regions are manually removed so that the remaining dataset consists of only nuclear regions. The entire step of preprocessing is shown in Figure 2.
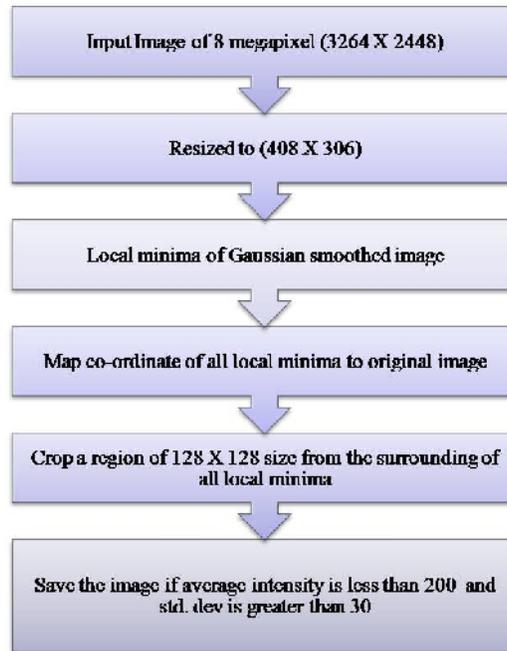
Figure 2. Steps of preprocessing stage

3.3. Segmentation

The identified nuclear regions are fed to the segmentation module. The staining of sputum cytology images generally does not give uniform color distribution. So identifying nuclear region based on one particular color may not give good segmentation result. The information content in RGB color space is highly correlated and to uncorrelate it, various color spaces is used. Some information may be discernable in one color space may not be quiet well distinguished in another color space. A novel method of combining various color spaces is used in the segmentation module. The technique is a generalization of the method used by Issac et al [31] where K-means clustering is done on LAB color space. In this method the input image is converted to LAB, LUV, HSV and xyY color spaces and various combination pair from each color space is taken and K-means/FCM clustering is performed. The 20 combination consists of LAB layer L-A, LAB layer A-B, LAB layer L-B, HSV layer H-S, HSV layer S-V, HSV layer H-V, xyY layer x-y, xyY layer y-Y, xyY layer x-Y , LUV layer L-U and HSV layer S only. The different combination gives a different result for the segmentation using K-means and FCM algorithm.

 The K-means and FCM algorithms give different labels for nuclei region in different images. So it is required to find the correct label for the nuclei region alone. The image cropping done in the

preprocessing stage ensures that nuclei region is present near the middle of the image. So 100 pixels at the center of the image is taken and the median value is found. The label of the median is considered as the label for the nucleus. Then morphological closing followed by opening is done to join small unconnected regions in the image using a 3 x 3 square structuring element. Then hole filling is done to remove holes inside nucleus region. From the resulting image the object with largest area is separated. This region gives the result of segmentation. This process is done on all the 20 results. Finally the best segmentation result has to be chosen. For this a weight value is given for each of the segmentation result. This is given by

$$ weight = \frac{area}{convex\ area}\ \frac{pixel\ in\ border}{total\ border} $$

(24)

The criterion for giving the first part is that most of the nuclear regions are of elliptical shape or shapes closer to that. So the ratio of area to convex area gives higher weight to convex shapes than concave shapes. The K-means and Fuzzy C-means algorithm are not guaranteed to find two separate regions sometimes only one region is returned i.e., the entire image itself giving maximum weight to such an image. So this condition has to be eliminated, for this a penalty term is added. The second term gives a penalty depending on the number of pixels that are clustered touches the boundary of the image. If no pixel is touching the border then the penalty is zero otherwise a positive value. If an erroneous clustering result comes to cover the entire image, the penalty value become 1 and so also the first term, giving 0 value effectively. Thus the image is eliminated from further analysis being the lowest weight value. From the 20 segmented results one with maximum weight is taken as the final segmented result. The various step of segmentation is shown in Figure 3.
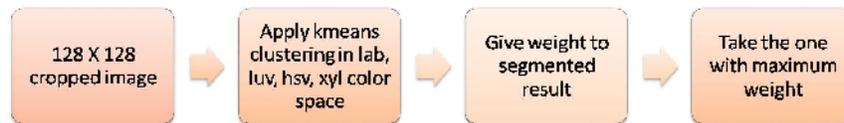
Figure 3. The steps of segmentation algorithm

4. RESULTS AND DISCUSSION

In this work we have used 1038 cell images which are cropped from the input sputum cytology sample image. The sputum cytology image is acquired using a trinocular microscope and is shown in the Figure 4. The cellular regions are cropped from the input image and are stored for further analysis. The cropped image from the input is shown in Figure 5. A sample cell image for benign and malignant is shown in Figure 6. The results of using various combinations of color spaces and clustering algorithm along with the color space, layer used along with weight value is shown in Figure 7. The input image along with the best segmentation result the algorithm has chosen using equation 25 is shown in Figure 8. Majority of the cell images were properly segmented. There are a few cases were the segmentation is not satisfactory due to various defects like non uniform absorption of stain by nuclear region and/or presence of mucus over cells as shown in Figure 9.
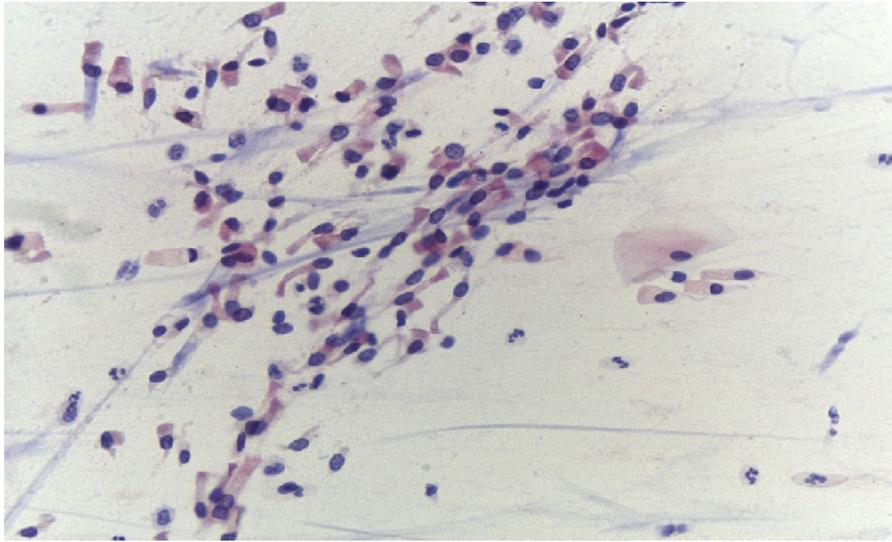
Figure 4. Input sputum cytology image provided by RCC, Thiruvananthapuram
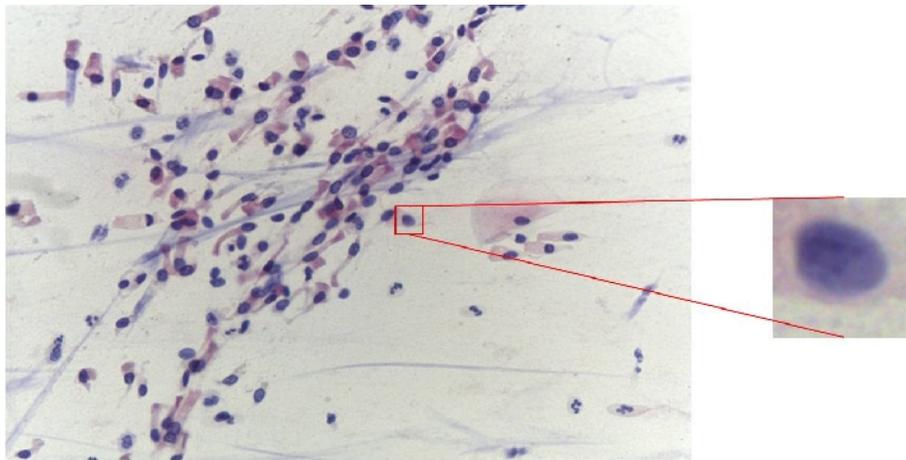
Figure 5. The cropped cell from the input image
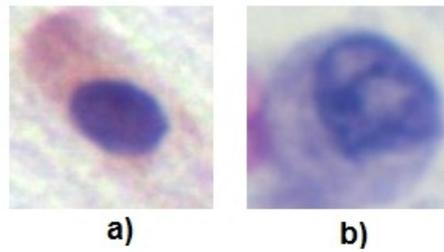
Figure 6. a) Benign cell and b) Malignant cell
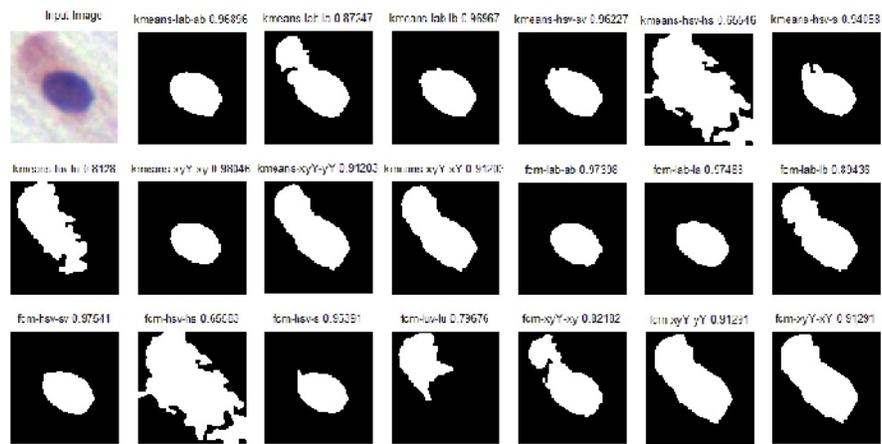
Figure 7. Input image along with 20 segmentation results
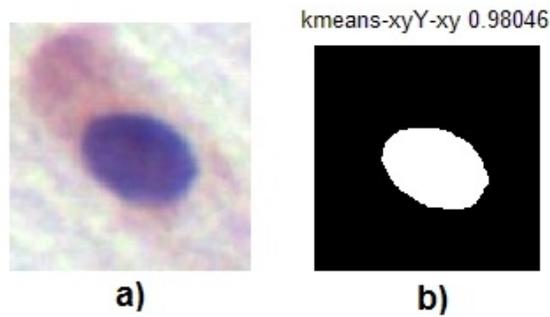
Figure 8. a) Input image, b) Best segmentation result
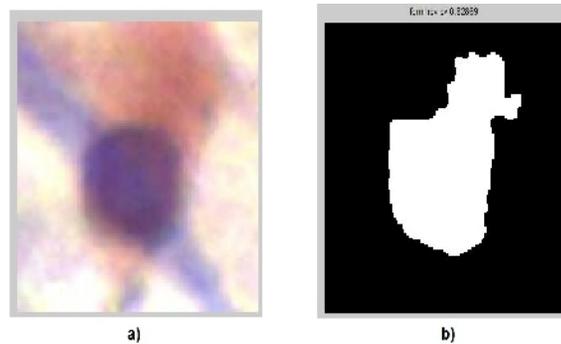
Figure 9. a) Input image and b) improper segmentation result

5. CONCLUSION

In this work, lung glandular cell nuclei detection and segmentation were done as the first step towards classification of slides as benign or malignant. In our method we used multiple color spaces and clustering algorithms such as K-means and Fuzzy C-means. The algorithm was tested on 1038 cell images. As a part of our future work we plan to include other methods so that the segmentation can be improved.

ACKNOWLEDGEMENTS

REFERENCES

[1]   R. F. Walker, P. Jackway, B. Lovell, & I. D. Longstaff, (1994)"Classification of cervical cell nuclei using morphological segmentation and textural feature extraction", in Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, pp 297-301.

[2]   J. P. Thiran & B. Macq, (1996) "Morphological feature extraction for the classification of digital images of cancerous tissues", IEEE Transactions on Biomedical Engineering, Vol. 43, pp 1011-1020

[3]   F. Schnorrenberg, C. S. Pattichis, K. C. Kyriacou, & C. N. Schizas, (1997) "Computer-aided detection of breast cancer nuclei", IEEE Transactions on Information Technology in Biomedicine, Vol. 1, pp 128-140.

[4]   H. Wu, J. Gil, & J. Barba, (1998)"Optimal segmentation of cell images", IEE Proceedings - Vision, Image, and Signal Processing, Vol. 145, No. 1, pp 50-56.

[5]   M. Sammouda, N. Niki, T. Niki, & N. Yamaguchi, (2000)"Analysis of color images of tissues derived from patients with adenocarcinoma of the lung", in IEEE Proceedings of 2000 International Conference on Image Processing, Vol. 1, pp192-195.

[6]   T. Tanaka, T. Joke, & T. Oka, (2001)"Cell nucleus segmentation of skin tumor using image processing", in Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society,Vol. 3, pp 2716- 2719.

[7] C. Wittke, J. Mayer, & F. Schweiggert, (2007)"On the classification of prostate carcinoma with methods from spatial statistics", IEEE Transactions on Information Technology in Biomedicine, Vol. 11, pp 406-414.

[8] V. Meas-Yedid, S. Tilie, & J. C. Olivo-Marin, (2002)"Color image segmentation based on markov random field clustering for histological image analysis", in Proceedings 16th International Conference on Pattern Recognition, Vol. 1, pp 796-799.

[9] G. Begelman, E. Gur, E. Rivlin, M. Rudzsky, & Z. Zalevsky, (2004)"Cell nuclei segmentation using fuzzy logic engine", in IEEE International Conference on Image Processing, Vol. 5, pp 2937- 2940.

[10] M. Hu, X. Ping, & Y. Ding, (2004)"Automated cell nucleus segmentation using improved snake", in International Conference on Image Processing, Vol. 4, pp 2737- 2740.

[11] P. Phukpattaranont & P. Boonyaphiphat, (2006)"Segmentation of cancer cells in microscopic images using neural network and mathematical morphology", in IEEE International Joint Conference SICE-ICASE, pp 2312–2315.

[12] B. Ko, M. Seo, & J. Nam, (2007)"Microscopic image segmentation for the clinical support system", in IEEE Computer Graphics, Imaging and Visualisation, pp 489-494.

[13] G. Palacios & J. R. Beltran, (2007)"Cell nuclei segmentation combining multiresolution analysis, clustering methods and colour spaces", in IEEE International Conference on Machine Vision and Image Processing, pp 91-97.

[14] F. Taher & R. Sammouda, (2011)"Lung cancer detection by using artificial neural network and fuzzy clustering methods" in IEEE GCC conference and Exhibition, pp 295-298.

[15] M. E. Plissiti, C. Nikou, & A. Charchanti, (2011)"Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering", IEEE Transactions on Information Technology in Biomedicine, Vol. 15, pp 233-241.

[16] W. Wang, J. A. Ozolek, D. Slepcev, A. B. Lee, C. Chen, & G. K. Rohde, (2011)"An optimal transportation approach for nuclear Structure-Based pathology", IEEE Transactions on Medical Imaging, Vol. 30, pp621-631.

[17] Z. Huang & D. Liu, (2007)"Segmentation of color image using EM algorithm in HSV color space", in IEEE International Conference on Information Acquisition, pp 316–319.

[18] T. Chen, Y. Chen, & S. Chien, (2008)"Fast image segmentation based on KMeans clustering with histograms in HSV color space", in IEEE 10th Workshop on Multimedia Signal Processing, pp 322-325.

[19] G. Zhao, S. Wang, T. Wang, & J. Chen, (2008)"HSV color space and face detection based objectionable image detecting", in Second International Conference on Future Generation Communication and Networking Symposia, Vol. 3, pp 107-110.

[20] Y. Lu, W. Gao, & J. Liu, (2010)"Color matching for colored fiber blends based on the fuzzy c-mean cluster in HSV color space", in Seventh IEEE International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 1, pp 452-455.

[21] V. Anari, P. Mahzouni, & R. Amirfattahi, (2010)"Automatic extraction of positive cells in pathology images of meningioma based on the maximal entropy principle and HSV color space", in 6th Iranian conference on Machine Vision and Image Processing, pp 1-4.

[22] M. H. Asmare, V. S. Asirvadam, & L. Iznita, (2009)"Color space selection for color image enhancement applications", in IEEE International Conference on Signal Acquisition and Processing, pp 208-212.

[23] H. Rami, M. Hamri, & L. Masmoudi, (2009)"Hybrid color space and support vector machines for classification", in IEEE International Conference on Multimedia Computing and Systems, pp 484-487.

[24] N. M. Kwok, Q. P. Ha, & G. Fang, (2009)"Effect of color space on color image segmentation", in 2nd IEEE International Congress on Image and Signal Processing, pp 1-5.

[25] Y. Pu, X. Wei, Y. Su, Z. Zhao, W. Qian, & D. Xu, (2010)"Implementation and research of multi-method color transfer algorithms in different color spaces", in IEEE 2010 International Conference on

Machine Vision and Human-Machine Interface , pp 713-717.

[26] Z. Yux, W. Yuhua, & L. Haibo, (2010)"A fuzzy c-means clustering based on hybrid color space", in IEEE International Conference on Electrical and Control Engineering, pp 4605-4607.

[27] D. Hang & L. Xi, (2010)"The color components' computing on different color spaces for image segmentation of hematocyte", in IEEE International Symposium on Intelligence Information Processing and Trusted Computing, pp 75–78.

[28] S. Chen & C. Liu, (2011)"Fast eye detection using different color spaces", in IEEE International Conference on Systems, Man, and Cybernetics, pp 521-526.

[29] L. Bradbury & J. W. Wan, (2010)"A spectral k-means approach to brightfield cell image segmentation", in Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 4748–4751.

[30] J. Macqueen, (1967)"Some methods of classification and analysis of multivariate observations", in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp 281–297.

[31] S. Issac Niwas, P. Palanisamy, R. Chibbar, & W. J. Zhang, (2011)"An expert support system for breast cancer diagnosis using color wavelet features", Journal of Medical Systems, pp 1-12.

[32] R. C. Gonzalez & R. E. Woods (2007) Digital Image Processing. Prentice Hall.

[33] A. Kuijper, (2004)"On detecting all saddle points in 2D images", Pattern Recognition Letters, Vol. 25, pp 1665-1672.