

Thresholding of Histopathological Images of Oral Mucosa for Identification of Precancerous Oral Submucous Fibrosis (OSF) Cells: A Novel Entropy based Approach

Saptarashmi Bandyopadhyay¹, Soumyadeep Basu², R. R. Paul³ and Ajoy Kumar Ray⁴

¹Department of Computer Science and Technology, IITEST, Shibpur, India

²Department of Physics, IIT Kharagpur, Kharagpur, India

³Department of Oral & Maxillo-facial Pathology, GNIDSR, Kolkata, India

⁴Department of Electronics and Electrical Communication Engineering, IIT Kharagpur, Kharagpur, India
saptarashmicse@gmail.com

Keywords: Entropy, Entropy-based Thresholding, Oral Submucous Fibrosis, Nuclear-Cytoplasmic Ratio

Abstract: The problem of early detection of Oral Submucous fibrosis (OSF) has received paramount importance in recent times. OSF is a chronic, irreversible and high risk pre-cancerous state of the oral mucosa. This kind of inflammatory and progressive fibrosis of the submucosal tissues is linked to oral cancers. This state results from chewing of areca nut which is prevalent in large parts of the Indian subcontinent. The current work presents an approach for the analysis of dysplastic epithelial cells from OSF, based on nuclear-cytoplasmic (N:C) ratio which is one of the most important morphological features to distinguish between normal and dysplastic epithelial cells. The proposed approach uses MATLAB to analyse the OSF biopsy images. This may help pathologists in identification of pre-cancer affected cells and in prevention and treatment of oral cancer. The methodology presented here can also be used for identification of epithelial atypia, an important light microscopic criteria that differentiates between normal and pre-malignant / malignant status of the oral mucosa.

1 INTRODUCTION

Oral submucous fibrosis (OSF) (Muthu et. al., 2012a, Muthu et. al., 2012b, Mahanta and Bora, 2012) is a chronic, potential pre-cancerous state of the oral mucosa, which is characterized by juxta-epithelial inflammatory reaction and progressive fibrosis of the submucosal tissues (lamina propria and deeper connective tissues). As the disease progresses, the mucosa becomes firm and rigid. As a result of which the victim is unable to open his/her mouth fully. The condition is linked to oral cancers caused by betel quid or areca nut chewing, a habit similar to tobacco chewing, practiced predominantly in Southeast Asia and India, dating back thousands of years. The incidence of oral leukoplakia, oral submucous fibrosis or squamous cell carcinoma are rapidly increasing in India and neighboring countries.

The oral sub-mucous fibrosis (OSF) is a slow progressive fibrotic disorder in oral mucosa with (OSFWD-Oral submucous fibrosis with dysplasia)

or without (OSFWT-Oral submucous fibrosis without dysplasia) epithelial dysplasia embedding features of chronic inflammation involving multi-level aberrations in the related tissues (Figure 1). The histological / light microscopic features of non dysplastic samples of OSF are confirmed by the occurrence of atrophic surface epithelium backed by hyalinized avascular sub-epithelial connective tissue, while dysplastic form of OSF are confirmed with the presence of cellular and nuclear pleomorphism of the surface epithelial cells, characterized by nuclear hyperchromatia, dyskeratosis, increased and abnormal mitosis etc. The cases of Oral squamous cell carcinoma (OSCC) are detected by the presence of dysplastic epithelium showing active invasion to the sub-epithelial connective tissue with concomitant formation of neoplastic islands.

The ratio of the size (i.e., volume) of the nucleus of a cell to the size of the cell cytoplasm is termed as the **nuclear-cytoplasmic ratio (N:C)**. The N:C ratio indicates the maturity of a cell, i.e., when the cell matures, the nucleus size generally decreases. In

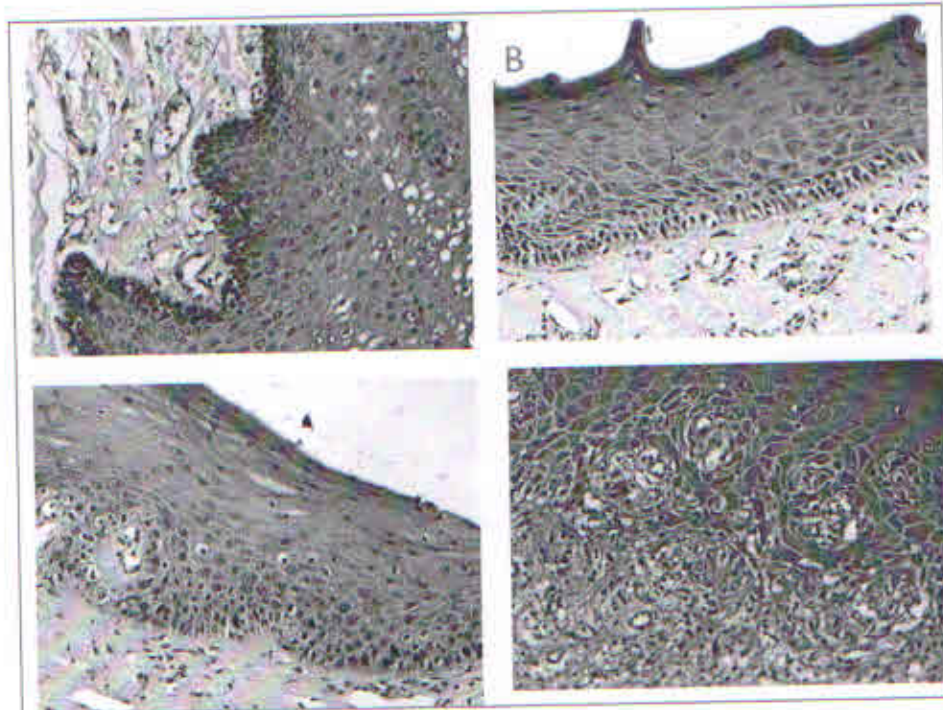


Figure 1: Haematoxylin and Eosin staining of the (A): Normal Oral Mucosa; (B) Oral Submucous fibrosis without dysplasia (OSFWT); (C) Oral Submucous fibrosis with dysplasia (OSFWD); (D) Oral squamous cell carcinoma (OSCC). (Source: patients biopsy from GNIDSR, Kolkata)

case of erythrocytes or leukocytes the nuclear-cytoplasmic ratio may reduce from as high as 4:1 to as low as 1:1 as the cells matures. On the other hand, precancerous dysplasia as well as malignant cells are commonly associated with increased N:C ratio. The classic malignant high N:C ratio cell is carcinoma of small cells.

Metqud et.al (2015) reported that there is a reduction of cytoplasmic diameter of exfoliated buccal mucosal cells in case of dysplastic changes in the lesions. We investigated whether cytoplasmic diameter, nuclear diameter and nucleus: cytoplasm ratio in exfoliative cytology can reliably identify potentially malignant lesions. A significant reduction in the mean cytoplasmic and nuclear diameter has been observed in the experimental groups compared to normal control.

Gao (1992) has conducted studies by interactive image analysis system (IBAS-II) on 48 specimens on the morphometry of the cells and the nuclei in the spinous cells of oral epithelium. A progressive reduction has been observed in the morphological features, e.g., area, perimeter and all kinds of diameter for the cells in the normal mucosa and also in the tissues affected in OSF, leukoplakia and

dysplasia to carcinoma ($P < 0.01$). No significant differences were observed in the dimensions of the nuclei among the groups ($P > 0.05$), except the minimum diameter. A progressive increase ($P < 0.01$) is observed in the nuclear cytoplasmic ratio. The shape factors seemed to be less helpful in identification. It is observed from the results that the decrease of the spinous cell area could reflect a malignant progression, and carcinoma can be distinguished from other lesions with nuclear-cytoplasmic ratio as an objective feature. It was also observed from the experimental results that OSF pathological grade lies between normal mucosa and mild epithelial dysplasia.

The N:C ratio for a cell can be identified as follows:

$$N:C \text{ ratio} = (\text{Area of nucleus}) / (\text{Area of the cytoplasm}) \quad (1)$$

That means if the area of the nucleus and cytoplasm are calculated in terms of the number of pixels occupying these regions in a cell then from Equation (1) one can calculate the N:C ratio easily. The nucleus of a normal cell consumes very less amount of area than the nucleus of a dysplastic cell. On the other hand, cytoplasm of a normal cell

consumes larger amount of area than that of a dysplastic cell. Thus the N:C ratio of a normal cell is less than N:C ratio of dysplastic cell. In the present work, the N:C ratio of nearly 100 individual cells obtained from the OSF cell biopsy images have been calculated. Previous work on analysis of malignant cervical cells based on N:C ratio using Pap Smear Images is reported in (Mahanta and Bora, 2012).

2 METHODOLOGY

The first step in the automated detection of oral submucous fibrosis from the light microscopic histopathological images of oral mucosa involves the segmentation of the light microscopic images of oral surface epithelium into individual cells. Figure 1 shows a certain cross-section of the image of surface epithelium, which is segmented into individual cells of the oral epithelium. Segmentation subdivides an image into its constituent regions or objects. Image segmentation algorithms use appropriate similarity measures of intensity values or color or texture. The nucleus and cytoplasm regions in a cell are next segmented by applying image thresholding techniques.

All the experiments on the OSF cell biopsy images have been carried out using the MATLAB tool. The colored cell images are converted to gray level images and thereafter to binary images so that the nucleus and cytoplasm areas in a cell are grossly identified and various image morphological operations can be performed on the binary image of a cell.

The process of image thresholding (Priyadarshini, 2014, Biswas and Ray, 2000, Zhang, 2011) involves binarization of a gray valued image using a single threshold for creating multiple number of segments using multiple number of thresholds. In this application we have segmented the histopathological images of oral mucosa and further segmented each cell into nucleus, identified as the black pixels and the rest part of the cell is identified as the cytoplasm. In the present work, various entropy based image thresholding methods have been experimented and evaluated.

Thresholding methods may be categorized as shape based, cluster based, entropy based, or spatial probability distribution based (Sakur and Sezgin, 2004) as follows:

- **Histogram shape-based** methods are multimodal in nature, where the curvatures, peaks and valleys of the smoothed histogram are analyzed.

- **Clustering-based** methods, where the samples in gray level are clustered in two parts as foreground and background, or alternately are modelled as a mixture of two Gaussians.

- **Entropy-based** methods lead to the development of algorithms that use the entropy of the background and foreground regions, the cross-entropy between the original and binarized image, etc.

- **Object Attribute-based** methods are concerned with searching a measure of similarity between the binarized and the gray level images, such as edge coincidence, fuzzy shape similarity, etc.

- **Spatial** methods make use of higher-order probability distribution and/or correlation between pixels.

- **Local** methods are used to adapt the threshold value on each pixel to the local image characteristics. Global area thresholding is the other spectrum in which the threshold value is adapted to clusters of gray level samples.

The entropy based methods are based on computing the entropies of the foreground and the background regions and also the cross entropy between the original image and the binarized image using a certain threshold. The threshold value is estimated as the one that optimizes the entropy measure.

Extensive studies (Zhang, 2011) have been conducted on thresholding algorithms for non-destructive testing (NDT) images as well for document images. It has been observed that entropy based image thresholding algorithms generally perform better for NDT images.

3 ENTROPY-BASED THRESHOLDING METHODS

The concept of entropy is used as a measure of diversity present in the image. The entropy for a grey level image is identified as follows: Each of the 256 grey levels is considered as an event in a grey level image. The probability of occurrence of each event in an image, i.e., for each grey level, is defined as the ratio of the total number of pixels of that grey value in the image to that of the total number of pixels in the image.

When a grey level image is converted into the binary image there is information transfer. Maximum information transfer occurs when the entropy of the thresholded image is maximized. The

intensity value in a grey level image for which the sum of the entropies for the foreground and background becomes maximum is identified as the threshold level which leads to the best binarisation (Sakur and Sezgin, 2004, Kapur et. al., 1985).

Among the several entropy based thresholding methods that can be applied are Shannon Entropy, Renyi Entropy, Yager Entropy, Pal's Entropy etc. Shannon Entropy is based on the distribution of gray levels in an image while Renyi Entropy is the generalization of Shannon Entropy. Both Yager Entropy and Pal's Entropy are fuzzy entropy in nature. In the present work, both Shannon and Renyi entropies and their variations have been considered.

3.1 Shannon Entropy

Specifically, Shannon entropy is the logarithm of the true diversity index with parameter equal to 1. The class of algorithms on entropy based thresholding methods exploits the entropy of the distribution of the gray levels in an image. The Shannon entropy is based on the weighted geometric mean of the proportional abundances of the types and it equals the logarithm of true diversity as calculated with parameter equal to 1:

$$H' = - \sum_{i=1}^R p_i \ln p_i \quad (2)$$

where p_i is the probability of occurrence of the i^{th} gray value.

3.2 Renyi Entropy

The Rényi entropy is a generalization of the Shannon entropy to values of diversity index parameter q other than unity. It can be expressed as:

$${}^q H = \frac{1}{1-q} \ln \left(\sum_{i=1}^R p_i^q \right) \quad (3)$$

In this method (Sakur and Sezgin, 2004, Kapur et. al., 1985) of entropy based image thresholding the foreground and background classes are considered as two different sources. The image is optimally thresholded as the sum of the foreground and background class entropies reach maximum. Thus, one has:

$$T_{opt} = \arg \max [H_f(T) + H_b(T)] \quad (4)$$

Thus the method aims to maximize the total entropy and the corresponding pixel value is identified as the threshold.

4 IMAGE THRESHOLDING IN THE PRESENT WORK

Normalized Shannon Entropy and Renyi Entropy values have been calculated for each pixel position in the grey level image with the real number entropy values are real numbers in the range [0,1]. The mean of the normalized Shannon or Renyi entropy values for all the image pixels can be defined as the mean entropy of the grey level image. The root mean square entropy of the normalized Shannon or Renyi entropy of all the pixels in the image gives an idea on the average divergence in the gray level image and can also be a basis for selecting the threshold.

The mean entropy of a grey level image is a real value in the range [0,1] and it must be maintained while it is converted to a binary image. Hence, the mean and rms entropy values of a grey level image are considered as the level in the *im2bw* function of MATLAB. All pixels in the image with intensity value above this level are converted to white while pixels with intensity value lower this level are converted to black.

Alternatively, complement mean or rms mean of the entropies of a grey level image which are real values in the range [0,1] can be defined as the mean entropy of the image. Hence, such entropy values of a grey level image are considered as the level in the *im2bw* function of MATLAB.

In addition, the log energy entropy values of a grey image find out the grey level that maximizes the entropy of the foreground and the background image. The specific grey level is then defined as the threshold grey level for conversion of the grey level image to binary one.

Thus, the following entropy measures have been directly used as the level for the grey level to binary image conversion:

1. Mean Renyi Entropy (MRE)
2. Complement Mean Renyi Entropy (CMRE)
3. Rms Renyi Entropy (RRE)
4. Complement rms Renyi Entropy (CRRE)
5. Log Energy entropy (LEE)
6. Rms Shannon Entropy (RSE)
7. Complement rms Shannon Entropy (CRSE)
8. Mean Shannon Entropy (MSE)
9. Complement Mean Shannon Entropy (CMSE)

5 N:C RATIO ALGORITHM

The algorithms for the N:C ratio is defined as:

1. Each individual cell is first segmented.

2. The background area is calculated as `area_background`.
3. The image is converted to grey scale image using `rgb2gray` function in MATLAB.
4. The contrast of the image is then adjusted using `imadjust` function.
5. Next Gaussian filter is applied on the image.
6. The `imhist` function of MATLAB is used to obtain the histogram of the image.
7. The grey scale image is converted to binary format using `im2bw` function of the image so that the morphological operation for feature extraction can be performed on the image.
8. Further experiments have been carried out for entropy based threshold for grey scale to binary image conversion.
9. The salt and pepper noise in the image is removed by applying Median filter on the output image.
10. The binary image is opened using `imopen` function of MATLAB which removes the thin protrusions. The `imopen` function also helps in finding if there are two separate nuclei in the cell which are slightly connected.
11. The `imcomplement` function of MATLAB is used to perform the complement operations.
12. Morphological erosion operation is performed using `imerode` function of MATLAB which is followed by Median filter of the image.
13. The holes are filled up using `imfill` function of MATLAB. Finally, an output image is obtained containing only the nucleus.
14. The area of the nucleus is calculated as `area_nucleus`.
15. The image of the cytoplasm is identified and the area is calculated as `area_cytoplasm`.
16. The N:C ratio is calculated as `area_nucleus / area_cytoplasm`.

6 EXPERIMENTS

The algorithm for identifying the N:C ratio is executed in three steps: calculation of nucleus (N) area, calculation of cytoplasm (C) area and finally, calculation of the N:C ratio. Each individual cell in the epithelium is segmented into nucleus and cytoplasm. The RGB coloured image is first converted to grey scale image. The grey scale image for an OSF cell consists of three areas: the nucleus, the cytoplasm and the background. The background cell image is generally white and its area is identified in terms of the number of white pixels as `area_background`.

The contrast of the image is then adjusted to get a uniform background so that the foreground image becomes more prominent. Next Gaussian low pass filter is applied on the image with parameters as `hsize` vector [5,5] and standard deviation 2. The histogram of the image is obtained in finding a suitable histogram based thresholding for grey scale to binary conversion of the image so that the morphological operations for feature extraction can be performed. Further experiments have been carried out for entropy based threshold for grey scale to binary image conversion. Table 1 shows an example original cell image and the threshold image. Finally median filter has been applied on the output image to remove the salt and pepper noise present.

Several morphological operations are then performed on the filtered image so that proper segmentation can be done. A flat disk-shaped structuring element is used for the morphological operation with radius 2. The binary image is opened which removes the thin protrusions. If there are two slightly connected separate nuclei in the cell this is also identified. The binary image is complemented so that all the pixels in the nucleus area are assigned the value of 0. Unwanted components like White Blood Cells (WBCs) and other cells create the main difficulty for proper segmentation. Morphological erosion operation is performed followed by Median filter of the image. Entire objects having less than 100 pixels are discarded to remove the WBCs present in the cell and the holes are filled up. Finally, an output image is obtained containing only the nucleus. The area of the nucleus is calculated in terms of the number of black pixels (pixels with grey value 0) and stored in the variable `area_nucleus`.






The area of cytoplasm in terms of the number of pixels in it is identified as total number of pixels in the cell image minus the number of background pixels (`area_background`) minus the number of pixels in the nucleus area (`area_nucleus`) and stored in the variable `area_cytoplasm`.

Finally, the N:C ratio is calculated as $N:C \text{ ratio} = \text{area_nucleus} / \text{area_cytoplasm}$.

Table 1: Original Image and Threshold Image

Original Image	Threshold Image
	

Table 2: N:C ratio computed by segmentation for different entropy based thresholding methods for different dysplastic and non dysplastic OSF cells (Source: The basal epithelial cells obtained from patients biopsy from GNIDSR, Kolkata).

Cell	N:C ratio computed by segmentation and thresholding using								
	1 (MRE)	2 (CMRE)	3 (RRE)	4 (CRRE)	5 (LEE)	6 (RSE)	7 (CRSE)	8 (SSE)	9 (CME)
	0.2965	0.6722	0.4913	0.5220	0.0309	0.8554	0.1283	0.7823	0.1870
	0.0785	0.4998	0.0980	0.2884	0.1565	0.2318	0.1157	0.1721	0.1492
	0.1646	0.9296	0.2437	0.6971	0.1478	0.6553	0.2700	0.4897	0.3537
	0.1833	0.3749	0.2799	0.4014	0.1039	0.6735	0.1484	0.5595	0.1922
	0.4060	0.3007	0.4808	0.2940	0.4378	0.6306	0.2740	0.5756	0.2819

7 EXPERIMENTAL RESULTS AND EVALUATION

Nearly 100 individual OSF surface epithelial cells have been analyzed - 50 of them are normal and 50 are dysplastic cells which are collected from OSF biopsy images. Table 2 lists some segmented cells including both normal and dysplastic cells, 5 results are shown in detail, among them 2 are normal (cell id 1,5) and rest 3 (cell id 2,3,4) are dysplastic OSF cells. Table 2 contains all 5 observations along with corresponding N:C ratio obtained on application of entropy based thresholding method. The method is identified by the method id mentioned as the

corresponding column heading. In column 1 of Table 2 the cell segmented from the biopsy image is shown. Preliminary observations reveal that mean Renyi (Method Id 1), rmsRenyi (Method Id 3), log Energy (Method Id 5), complement mean Shannon (Method Id 7) and Complement rms Shannon entropy (Method Id 9) metrics produce good results. The rmsRenyi (Method Id 3) produces conclusive results as cells having N:C ratio less than equal to 0.3 can be considered as dysplastic OSF cells while cells having N:C ratio more than 0.3 can be considered as normal cells. The Log Energy based method also produces conclusive results as cells having N:C ratio less than equal to 0.1 can be considered as normal while cells having N:C ratio

more than 0.1 can be considered as dysplastic OSF cells. The evaluation results on 100 cells – 50 normal and 50 dysplastic category cells have been shown in Table 3. Since, the rms Renyi Entropy was giving the best results, we identified a cell as normal or dysplastic based on the decision obtained from the rms Renyi Entropy method. The accuracy of cell classification using the simple N:C ratio feature has been found to be 88%. Detailed experiments need to be considered with more number of normal and dysplastic cells.

Table 3: Evaluation Results.

Identified as Cell Category	Normal	Dysplastic
Normal - 50	43	7
Dysplastic	5	45

8 CONCLUSION

In the present work, N:C ratios have been obtained for around 100 segmented cells belonging to either normal or OSF cells with dysplasia. It has been observed that normal cells have N:C ratio within 0.03 while dysplastic cells have N:C ratio greater than 0.03 based on the rms Renyi Entropy method. But, N:C ratio alone is not sufficient to identify dysplastic cells. Moreover, the area of the cytoplasm has been computed after thresholding the individual basal cells from epithelium. The basal epithelial cells have been obtained from patients biopsy from Guru Nanak Institute of Dental Sciences and Research (GNIDSR), Kolkata. While some of these cells are normal, others are affected with dysplasia. With the introduction of more number of features, the overall accuracy of cell classification will surely be enhanced.

REFERENCES

- M. Muthu, Rama Krishnan, Chandan Chakraborty, Ranjan Rashmi Paul, Ajoy Kumar Ray, 2012a. Hybrid segmentation, characterization and classification of basal cell nuclei from histopathological images of normal oral mucosa and oral submucous fibrosis. *Expert Syst. Appl.* vol. 39(1), pp. 1062-1077.
- M. Muthu, Rama Krishnan, Pratik Shah, Chandan Chakraborty, Ajoy Kumar Ray, 2012b. Statistical Analysis of Textural Features for Improved Classification of Oral Histopathological Images. *Journal of Medical Systems*, vol. 36(2), pp. 865-881.
- Metgud R, Gupta K, Prasad U, Gupta J, 2015. Cytomorphometric analysis of oral submucous fibrosis and leukoplakia using methyl green-pyronin Y, Feulgen staining and exfoliative brush cytology. *Biotech Histochem*, Vol. 90(1), pp. 8-13.
- Mahanta Lipi B and Bora Kangkana, 2012. Analysis of Malignant Cervical Cells based on N/C ratio using Pap Smear Images. *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, Issue 11, pp. 342-346.
- Priyadarshini B., 2014. Classification System for Oral submucous Grading – A Review. *International Journal of Science and Research*, ISSN (online) 2319-7064, vol. 3, Issue 3, pp. 740-744.
- Jawahar, Biswas and Ray, 2000. Analysis of Fuzzy Thresholding Schemes. *Pattern Recognition*, vol. 33, Issue 8, pp. 1339-1349.
- Zhang, Y., 2011. Optimal multi-level Thresholding based on Maximum Tsallis Entropy via an Artificial Bee Colony Approach. *Entropy*, vol. 13, Issue 4, pp. 841-859.
- Bulent Sakur and Mehmet Sezgin, 2004. Survey Over Image Thresholding Techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, vol. 13, Issue 1, pp. 146-165.
- J. N. Kapur, P. K. Sahoo, A.K.C. Wong, 1985. A New Method for Gray-Level Picture Thresholding Using the Entropy of the Histogram. *Graphical Models and Image Processing*, vol. 29, pp. 273-285.
- Gao S 1992. Cell morphometric analysis in oral submucous fibrosis, leukoplakia and squamous cell carcinoma. [Article in Chinese]. *Chinese Journal of Stomatology*, vol. 27(3), pp. 145-147.