Journal of Biology and Today's World

Journal home page: http://journals.lexispublisher.com/jbtw

Received: 26 October 2017 • Accepted: 08 January 2018

Research

doi:10.15412/J.JBTW.01070106

Clustering Techniques on Pap-smear Images for the Detection of Cervical Cancer

Mithlesh Arya, Namita Mittal, Girdhari Singh*

Department of Computer Science and engineering, Malaviya National Institute of Technology, Jaipur, Rajasthan, India

*Correspondence should be addressed to Girdhari Singh Department of Computer Science and engineering, Malaviya National Institute of Technology, Jaipur, Rajasthan, India; Tel: +91xxxxx; Fax: +91xxxx; Email: girdharisingh@rediffmail.com.

ABSTRACT

A Pap smear test is the most efficient and prominent method for the detection of dysplasia in cervical cells. Pap smear is time-consuming and sometimes it is an erroneous method. Computer-assisted screening can be widely used for cervical cancer diagnosis and treatment. Most of the existing approaches do not give good performance on real images due to poor staining, dye, blood and inflammatory cells. In our proposed approach, we are extracting nucleus only from the Pap smear images. For segmentation Laplacian of Gaussian (LOG) filter and morphological operations has used for edge detection. In the classification phase, two clustering techniques K-means and Fuzzy c-means (FCM) has been used using Principle Component Analysis (PCA). The classification of Pap smear images is based on the Bethesda System. The approach has performed on a dataset obtained from pathologic lab containing 40 Pap smear images with 500 cells. Performance evaluation has done using Purity and Jaccard Index (JI). The purity of K-means is 0.815 and for FCM it's 0.875.

Keywords: Pap-smear, Cervical Cancer, Clustering Techniques.

Journal of Biology and Today's World is published by *Lexis Publisher*; Journal p-ISSN 2476-5376; Journal e-ISSN 2322-3308.

1. INTRODUCTION

In India, cervical cancer is the most common cause of female mortality. In the year 2015 total number of 1,22 500 cases of cervical cancer were detected and out of them, 67,400 lost their life (1). According to the latest census, female population aged 15 years and more is 432 million in India and this is the number which is at risk of acquiring cervical cancer. The peak incidence of detecting cancer is 54-59 years of age. There are many reasons for developing cervical cancer like lack of awareness, early marriage, prolonged use of contraceptive pills, multiple partners, poor hygiene, low immunity and much more (2). Infection of Human Papilloma Virus (HPV) is strongly associated with cervical cancer. Vaccinations against many strains of HPV including HPV 16 and 18 are available in the market, but due to lack of awareness these preventive measures are not in very much use. Although the government of India started many awareness programs but it is in very early phase. Cervical cancer is the cancer of female reproductive organ uterus. The cervix is lowermost part of the uterus. Recently trend of cancer is on down trends due to early detection. For the detection of cancer, many tests are available but Pap's Smear (Papanicolaou test) test is most commonly used screening test. Pap's Smear Test was first demonstrated by the scientist George Papanicolaou in 1940 (3). Pap test helps in detecting precancerous changes in the cervical cells. Pap's Smear test is of 2 types 1) Conventional 2) Thin preparation. Both techniques are different in the way the sample is obtained. In Pap test, cells are scraped from the cervical cell lining and then cells are spread over a glass slide. Cells obtained are mostly from the superficial layer. The cell-laden slides are then stained with a dye called Methylene blue and allowed to dry. A stained slide containing cervical cells and other cells are examined under a microscope. In the cervix there are many different types of cells are present like Squamous Epithelial Cells and Columnar Cells. Squamous epithelial is further divided according to their morphological features into 4 layers namely basal, parabasal, intermediate and superficial. Normally the nucleus to cytoplasm ratio is 1:4 to 1:6 but in precancerous cells, the ratio gets disturbed that is the nucleus size becomes many times of that normal nucleus size. Limitations of this procedure are that it is very time consuming as well as a lot of experience is required to classify the cells according to their morphological findings. Many automated and semi-automated system are proposed using different segmentation and classification methods.

Most of the existing algorithms have worked on DTU/HERLEV dataset (4), which is a single cell image. In Real dataset have overlapped and poorly stained cell with a blood clot and inflammatory cells. Therefore, we are developing a more advanced methodology for the screening of the Pap's Smear images which is more accurate on real images. Abnormal cervical cells are called as dysplasia which is also cervical intraepithelial Neoplasia (CIN) (5). CIN is further classified into three categories namely:

1. CIN 1: Mild Dysplasia
2. CIN 2: Moderate Dysplasia
3. CIN 3: Severe Dysplasia

Above mentioned method of classification is a conventional one. However, we are using a system of classification which is approved and recently been updated by the association of pathologists that is The Bethesda System (TBS) (6). The Bethesda system (TBS) is used for reporting of cancerous and pre-cancerous stages for Pap smear results. According to TBS cervical dysplasia is categorized into three levels:

1. Normal
2. Low-grade Squamous Intraepithelial Lessons (LSIL)
3. High-grade Squamous Intraepithelial Lessons (HSIL).

1.1. Related work

There have been many studies on cervical cancer. Many automated and semi-automated system have been proposed. All proposed methods can be categories into four factors:

Consideration of single cell or multiple cells
Type of segmentation algorithms used
Type of features extracted
Type of classification method used

For the segmentation of ROI, many techniques have proposed. The paper (7) proposed a generic image processing method to segment nucleus and cytoplasm by using Gaussian Mixture Model (GMM), each pixel is assigned to a class based on its weight associated with a component in a mixture of distributions. The parameters for Gaussian distribution have been calculated using Expectation-Maximization (EM) algorithm. In paper (8) thresholding value and morphological closing methods have been used for segmentation, then by using different line masks like horizontal, vertical, +45 and -45 nucleus boundary have been defined. In paper (9) GMM and EM algorithms with K-means clustering has been used for segmentation of the image in background, nucleus and cytoplasm regions. In paper (10) J 1.44C has been used for segmentation and preprocessing. J image is an application for image processing. In paper (11) watershed algorithm has been used for segmentation to identify the area of background, nucleus, and cytoplasm. In paper (12) multiple morphological operations have been used for segmentation to get the nucleus from the Pap image and Gaussian function to extract the ROI. Literature has been categorized based on different features used to show change. In paper (8) basic features like area, perimeter, pixel minimum value, pixel maximum value, standard deviation and mean has been extracted. In paper (9) multiple features like the area of nucleus and cytoplasm in the form of the pixel, the ratio of N/C, the brightness of nucleus and cytoplasm, the perimeter of the cell, the major and minor axis of cell, nucleus location in the cell has been extracted. In paper (13) single and multiple features comparative analysis has done and showing the importance of each feature for classification and better accuracy. In paper (11) Gray Level Co-occurrence Matrix (GLCM), Haralick, Gradient and Tamura based features have been extracted. Lots of features have been extracted but it has not been mentioned that which one has given the better results. In paper (12) 34 standards GLCM features have been extracted and Gaussian function have used to extract the prominent feature. Classification is the last but most important stage in designing any Decision Support System. Literature reflects the cell classification mainly focus in single cell study into normal and abnormal classes. Smear level classification is comparatively difficult than the single cell classification. Standard methods for classification like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN) and Decision Tree have been used. In paper (14) single cell classification into two classes using SVM has been done. The performance of SVM has compared with ANN and KNN. In paper (15) decision tree has been used for the single cell classification into four classes. In paper (16) minimum distance classifier and KNN has been used for single cell classification into two classes. In the traditional method of Pap smear, there are 100 to 10000 cells. Cells of different shapes, size and varieties were present. Cells like inflammatory cells, Red Blood Cells (RBC), cell debris were present in Pap smear slide. In cervical cancer, cancerous cells go under many changes including shape, size, color and texture and these morphological changes are considered as features for the classification of cells. Most of the work has done on a single cell and on multiple cells but in our dataset we use actual smear images containing cell debris. So we focused exclusively on the nucleus of the cell.

1.2. Data collection

In our study, we are capturing images using the high-resolution digital camera (OLYMPUS C 5060) which is mounted on a microscope (OLYMPUS BX 51) and images are stored in a digital format with tiff extension. Magnification of images can be done at various scales like 10x, 20x, 40x and 100x. We are using 40 x magnifications. Images obtained are displayed at a resolution of

2560x1920 with 24 bits color depth. In our study 40, Pap smear images have been collected which contain at least 500 cells. For the validation of our work, DTU/ HERLEV Pap smear benchmark dataset (4) which has been collected by the department of pathology at HERLEV university hospital and the department of automation at the technical university of Denmark has also been used. The dataset consists of 917 images which are classified into seven classes. The first 3 classes correspond to normal cells and remaining 4 classes correspond to abnormal cells. Cell distribution is mentioned below in Table 1 in the data set.

In the proposed work clustering techniques has used for cervical cancer detection using Pap smear images. The method is presented in the Figure 1 given below. Our proposed work has divided into 4 phases in 1st phase data set has prepared using Pap smear slides. In phase 2 preprocessing and segmentation has done for the extraction of the reason of interest. Phase 3 for feature extraction with fewer features for better performance. We have extracted the shape and size of the nucleus. In the final phase, we classified our data into three classes according to Bethesda system using K-means and FCM clustering technique. This same method has applied on both the datasets.

2. MATERIALS AND METHODS

Table 1. DTU/Herlev dataset Description

Cell type	**No. of Cells**
Superficial Squamous Epithelial	97
Intermediate Squamous Epithelial	70
Columnar Epithelial	74
Mild Dysplasia	182
Moderate Dysplasia	146
Severe Dysplasia	197
Carcinoma In Situ	150

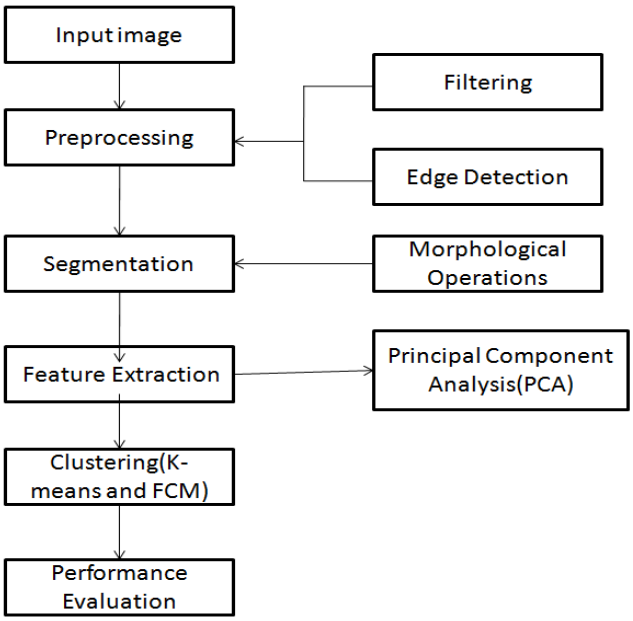


Figure 1. Steps of proposed method

2.1. Preprocessing and Segmentation

Preprocessing and segmentation are the only tools to extract the reason of interest from the image. Our focus in the proposed work is to extract similar features from any image through segmentation technique. For obtaining efficient segmentation results we are using following steps. RGB color image converted into a gray image. The sharp function has used to enhance the edges of the cells and to reduce the darkness of background bright function used. To remove noise and background debris Laplacian of Gaussian (LOG) filter has used. LOG basically used for edge detection on a smooth image. The first image has smoothened by Gaussian filter than Laplacian edge detector used. The Laplacian L(x, y) of an image with pixel intensity value I (x, y) worked as;

$$L(x, y) = \partial_2 I/\partial x^2 + \partial_2 I/\partial y^2 \quad (1)$$

To extract the nucleus only from the gray image threshold function has used. Otsu method (16) has not worked well on our images. We have applied multiple threshold values. For normal image threshold value vary from 90 to 100 and for cancerous image threshold value vary from 50-60 only. After threshold canny edge detection and gradient function has used. To find the exact boundary of nucleus, morphological operations (17) has used. The segmented image has used dilation function to add the pixel in the boundary by using the corresponding value of neighbor

pixels. Erosion function has removed the extra pixel from the boundary. In the last step, the mask has subtracted from the actual image to get the exact nucleus. Figure 2 and Figure 3 show the preprocessed and segmented image of normal cell and abnormal (HSIL) cell. As we can see in Figure 2 and Figure 3 the size and shape of the nucleus has changed. Normal cells have small and round nucleus but when the normal cell is converted into abnormal cell its nucleus size is increased and shape become oval or elliptical.

2.2. Feature Extraction

Feature extraction is a process for transferring most relevant information from the original data set into a low dimensional space. In our work, the feature extraction is applied for converting microscopic images into quantitative and parametric values. Segmentation image gives a number of the nucleus in the smear image. We have extracted these six features (18) of the nucleus for further classification.

1. Area of nucleus in the term of pixel (A)
2. Perimeter of nucleus in the term of pixel (P)
3. Compactness of nucleus $C = P^2/A$
4. Major axis of nucleus (L)
5. Minor axis of nucleus (D)
6. Ratio of minor and major axis of nucleus R= D/L
7. Eccentricity $E = ((L^2-D^2)/D^2)^{1/2}$

Eccentricity (19) show the change in the shape of the nucleus. If the cell is normal then its nucleus is round in shape and the value of eccentricity is zero (near).

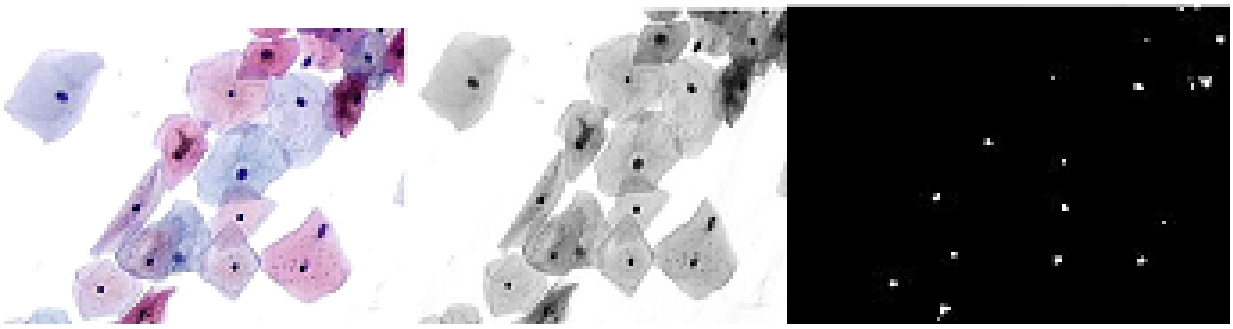


Figure 2. a) Pap smear of normal cell, b) Preprocessed and enhanced image, c) Final image with number of nucleus

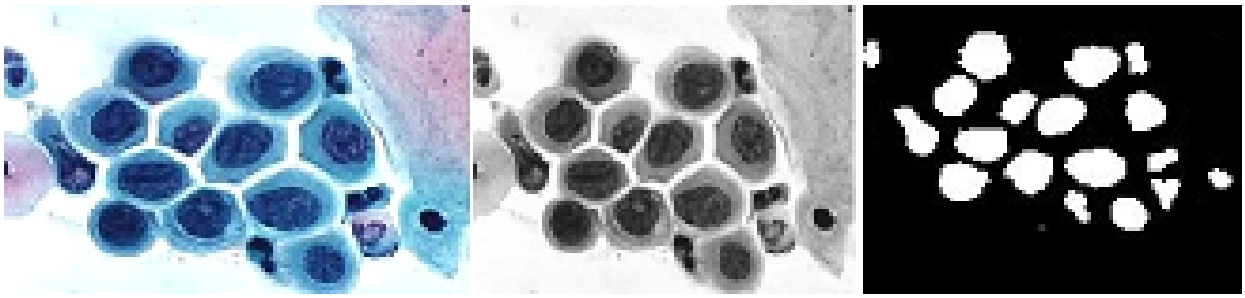


Figure 3. a) Pap smear image with abnormal cell, b) Enhanced and preprocessed image, c) Segmented image with Final image with number of nucleus

But if the cell is abnormal its eccentricity value has variation. We have calculated the individual value of all features for each nucleus in the single smear image. After that, calculate the mean value for the single image. Table 2 and Table 3, shows the feature values for normal cell and abnormal cell. Table 2 shows that the value of area varies from 55 to 208. Eccentricity value is near to zero. But, Table 3 shows the area value varies from 300 to 2000 and an eccentricity value is more than zero. In our dataset, we have 40 Pap smear images, 12 normal images, 14 LSIL images and 14 HSIL images. Herlev dataset has 241 normal images, 328 LSIL images, and 347 HSIL images.

Table 2. Area, Perimeter and Eccentricity of normal cell

Area	*Perimeter*	**Eccentricity**
68.8235	22.5312	0.9093
55.5499	17.4786	0.8596
184.5385	42.3876	0.8957
133.3234	36.6173	0.9818
208.0571	42.1356	0.8369

Table 3. Area, Perimeter and Eccentricity of abnormal cell

Area	*Perimeter*	**Eccentricity**
383.6500	71.7851	1.6797
617.1207	96.9537	1.4972
1892.4619	153.6309	1.7991
2246.5263	271.1046	2.9886
3147.4023	298.3192	1.9548

2.3. Classification

Usually, the classification techniques are based on shape, texture and color feature. However, we have classified on the basis of the shape of nucleus only. Cervical dysplasia is categorized according to Bethesda system into three classes: Normal, LSIL, and HSIL. Low-Grade Squamous Intraepithelial Lessons (LSIL) is the combination of CIN1 and CIN2 and High-Grade Intraepithelial Lessons (HSIL) is the combination of CIN3 and Carcinoma in Situ. In classification phase, we are using two clustering techniques. Three clusters have been made through K-means and Fuzzy C-means methods. The prominent features are selected by applying PCA on extracted features.

3. RESULTS AND EVALUATION

K-means and FCM techniques used for clustering. Three clusters have been made for Normal, LSIL, and HSIL cells. These external criteria have used for quality of clustering. Purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing. For good clustering, purity value should be one.

$$Purity(\alpha,C) = 1/N\sum_k \max(j)|wk \cap cj| \quad (2)$$

Where, α=(w1,w2,w3….wk) is the set of clusters and C=(c1,c2,c3…..cj) is the set of classes.

Jaccard Index (JI), measures the similarity between two clusters. JI value varies between 0 and 1.

$$JI = TP / (TP + FP + FN) \quad (3)$$

Where, TP is true positive values, FP is false positive values and FN is false negative values. Confusion matrixes of K-means and FCM have shown in Table 4 and Table 5. Table 6 and Table 7 show the confusion matrix of K-means and FCM using PCA.

Table 4. Confusion matrix of K-means

12	0	0	Normal
2	12	0	LSIL
0	3	11	HSIL

Table 5. Confusion matrix of FCM

12	0	0	Normal
2	12	0	LSIL
0	3	11	HSIL

Table 6. Confusion matrix of K-means with PCA

12	0	0	Normal
3	11	0	LSIL
0	2	12	HSIL

Table 7. Confusion matrix of FCM with PCA

12	0	0	Normal
3	11	0	LSIL
0	2	12	HSIL

Four prominent features have been selected out of seven features using PCA. The values of seven features are 74.7188, 17.8757, 5.3276, 1.8830, 0.1490, 0.0302 and 0.0156.

1. Purity of K-mean = 0.815
2. Purity of FCM = 0.875

3. JI of K-mean = 0.935
4. JI of FCM = 0.911

Table 8 shows the comparative performance of K-means and FCM with PCA and without PCA.

Table 8. Performance of K-means and FCM

Clustering Techniques/ Performance	*Purity*	**Jaccard Index**
K-means	0.815	0.935
FCM	0.875	0.911
K-means with PCA	0.875	0.881
FCM with PCA	0.850	0.911

Using PCA Purity performance of K-means has increased but the performance of FCM has decreased. Best result has obtained by K-means with PCA because the purity is high and the difference between clusters is also greater.

1. Purity of K-means (with PCA) = 0.875
2. Purity of FCM (with PCA) = 0.85
3. JI of K-mean (with PCA) = 0.881
4. JI of FCM (with PCA) = 0.911

4. CONCLUSION

The proposed method in this study for the detection of cervical cancer cells has given good results. In segmentation, LOG filter has worked better for removal of noise and debris. Two clustering techniques K-means and FCM have been used for the clustering of cells into three classes: normal, LSIL and HSIL. Purity and JI obtained for K- means 0.815 and 0.935 respectively. On the other hand, for FCM it is 0.875 and 0.911 respectively. One more factor PCA has been used for better performance but the results by using this factor are not satisfactory as the number of features in our study were less. In our future work, we will try to extract features of cytoplasm and nucleus from the cells so that we will get more features and hence PCA factor will give better results.

ACKNOWLEDGMENT

We would like to thanks Dr. Archana Pareek and Dr. Mukesh Rathore for their help in collection of Pap smear slide and generating dataset with findings.

FUNDING/SUPPORT

There has been no financial support for this work.

AUTHORS CONTRIBUTION

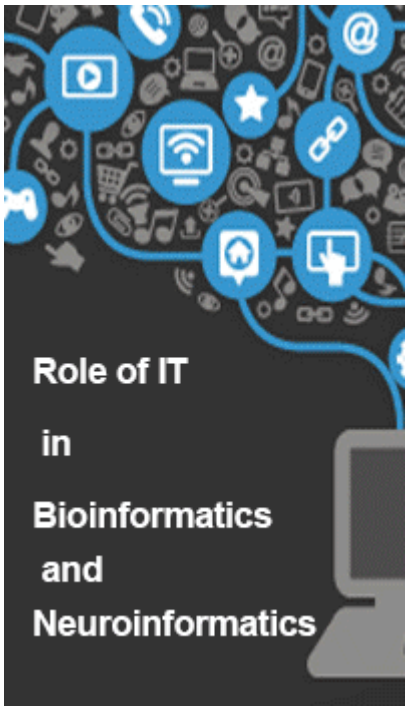
This work was carried out in collaboration with all the authors. Dr. Namita Mittal and Dr. Girdhari Singh contributed to the writing of the manuscript. Mithlesh Arya worked on dataset and experimental section.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this paper.

REFERENCES

1. Sreedevi A, Javed R, Dinesh A. Epidemiology of cervical cancer with special focus on India. Int J Womens Health. 2015;7:405-14.
2. Nordqvist C. Cervical Cancer: Causes, Symptoms, and Treatments. Medical News Today. 2015.
3. Tan SY, Tatsumura Y. George Papanicolaou (1883–1962): discoverer of the Pap smear. Singapore medical journal. 2015;56(10):586.
4. Sukumar P, Gnanamurthy R. Computer aided detection of cervical cancer using PAP smear images based on hybrid classifier. International Journal of Applied Engineering Research, Research India Publications. 2015;10(8):21021-32.
5. Nobbenhuis MA, Walboomers JM, Helmerhorst TJ, Rozendaal L, Remmink AJ, Risse EK, et al. Relation of human papilloma virus status to cervical lesions and consequences for cervical-cancer screening: a prospective study. The Lancet. 1999;354(9172):20-5.
6. Solomon D, Davey D, Kurman R, Moriarty A, O'connor D, Prey M, et al. The 2001 Bethesda System: terminology for reporting results of cervical cytology. Jama. 2002;287(16):2114-9.
7. Kale A, Aksoy S, editors. Segmentation of cervical cell images. Proceedings of the 2010 20th International Conference on Pattern Recognition; 2010: IEEE Computer Society.
8. Athinarayanan S, Srinath M, Kavitha R. Computer aided diagnosis for detection and stage identification of cervical cancer by using pap smear screening test images. ictact Journal on Image & Video Processing. 2016;6(4).
9. Lakshmi GK, Krishnaveni K, editors. Multiple feature extraction from cervical cytology images by gaussian mixture model. Computing and Communication Technologies (WCCCT), 2014 World Congress on; 2014: IEEE.
10. Divya Rani M, Narasimha A, Harendra Kumar M, Sheela S. Evaluation of Pre-Malignant and Malignant Lesions in Cervico Vaginal (Pap) Smears by Nuclear Morphometry. Journal of clinical and diagnostic research: JCDR. 2014;8(11):FC16.
11. Kenny SPK, Allwin S. Creating a Optimal Set of Textural Features for Cervical Cancer Lesions Using Hierarchal Clustering Technique.
12. Kenny SPK, Allwin S. Determining Optimal Textural Features for Cervical Cancer lesions using the Gaussian Function. jiP.1:1.
13. Kenny SPK, Victor S. A comparative analysis of single and combination feature extraction techniques for detecting cervical cancer lesions. ictact Journal on Image and Video Processing. 2016;6(3):1167-73.
14. Athinarayanan S, Srinath M. Classification of cervical cancer cells in PAP smear screening test. ICTACT Journal on Image and Video Processing. 2016;6(4):1234-8.
15. Paul PR, Bhowmik MK, Bhattacharjee D, editors. Automated cervical cancer detection using Pap smear images. Proceedings of Fourth International Conference on Soft Computing for Problem Solving; 2015: Springer.
16. Otsu N. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics. 1979;9(1):62-6.
17. Soille P. Morphological image analysis: principles and applications: Springer Science & Business Media; 2013.
18. Chankong T, Theera-Umpon N, Auephanwiriyakul S. Automatic cervical cell segmentation and classification in Pap smears. Computer methods and programs in biomedicine. 2014;113(2):539-56.
19. Mahanta LB, Nath DC, Nath CK. Cervix cancer diagnosis from pap smear images using structure based segmentation and shape analysis. Journal of Emerging Trends in Computing and Information Sciences. 2012;3(2):245-9.



Special Issue on:

'Role of IT in Bioinformatics and Neuro-informatics'

Guest Editors:

Prof. Deepshikha Bhargava
Amity University Rajasthan – India
dbhargava1@jpr.amity.edu

Dr. Ramesh C. Poonia
Amity University Rajasthan – India
rameshcpoonia@gmail.com

Dr. Swapnesh Taterh
Amity University Rajasthan – India
staterh@jpr.amity.edu

This paper is published as one of the selected papers that were presented at:
International Conference on Smart Computing and Communication (SmartTech-2017).