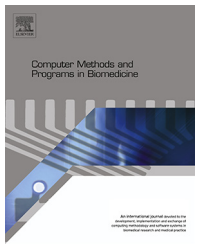
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 138 (2017) 31–47

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Automated classification of Pap smear images to detect cervical dysplasia

Kangkana Bora [a,*], *Manish Chowdhury* [b], *Lipi B. Mahanta* [a], *Malay Kumar Kundu* [c], *Anup Kumar Das* [d]

[a] *Department of Centre for Computational and Numerical Sciences, Institute of Advanced Study in Science and Technology, Guwahati 781035, Assam, India*
[b] *KTH, School of Technology and Health, Hälsovägen 11c, SE-141 57 Huddinge, Stockholm, Sweden*
[c] *Department of Machine Intelligence Unit, Indian Statistical Institute, 203 B.T.Road, Kolkata 700108, India*
[d] *Ayursundra Healthcare Pvt. Ltd, DMB Plaza, Lachit Nagar, Guwahati 781007, Assam, India*

ARTICLE INFO

Article history:
Received 12 April 2016
Received in revised form
19 August 2016
Accepted 4 October 2016

Keywords:
Pap smear
MSER
Ripplet transform
Ensemble classification

ABSTRACT

Background and objectives: The present study proposes an intelligent system for automatic categorization of Pap smear images to detect cervical dysplasia, which has been an open problem ongoing for last five decades.
Methods: The classification technique is based on shape, texture and color features. It classifies the cervical dysplasia into two-level (normal and abnormal) and three-level (Negative for Intraepithelial Lesion or Malignancy, Low-grade Squamous Intraepithelial Lesion and High-grade Squamous Intraepithelial Lesion) classes reflecting the established Bethesda system of classification used for diagnosis of cancerous or precancerous lesion of cervix. The system is evaluated on two generated databases obtained from two diagnostic centers, one containing 1610 single cervical cells and the other 1320 complete smear level images. The main objective of this database generation is to categorize the images according to the Bethesda system of classification both of which require lots of training and expertise. The system is also trained and tested on the benchmark Herlev University database which is publicly available. In this contribution a new segmentation technique has also been proposed for extracting shape features. Ripplet Type I transform, Histogram first order statistics and Gray Level Co-occurrence Matrix have been used for color and texture features respectively. To improve classification results, ensemble method is used, which integrates the decision of three classifiers. Assessments are performed using 5 fold cross validation.
Results: Extended experiments reveal that the proposed system can successfully classify Pap smear images performing significantly better when compared with other existing methods.
Conclusion: This type of automated cancer classifier will be of particular help in early detection of cancer.

* *Corresponding author.* Department of Centre for Computational and Numerical Sciences, Institute of Advanced Study in Science and Technology, Guwahati 781035, Assam, India. Fax: 0361-2273063.
E-mail address: kangkana.bora89@gmail.com (K. Bora).

http://dx.doi.org/10.1016/j.cmpb.2016.10.001
0169-2607/

1. Introduction

Cervical cancer starts in the cells lining the cervix (the lower part of the uterus). Normal cells of cervix gradually develop pre-cancerous changes first and then develop into cancer. Cervical cancer is the second most prevalent cancer after breast cancer among women, which is more prevalent in developing countries. It is stated that cervical cancer can normally be effectively treated if it is detected at an early stage [1]. Papanicolaou test (abbreviated as Pap test) is a method of cervical screening used to detect potentially pre-cancerous and cancerous process in the cervix. Pap test has made cervical cancer one of the most preventable cancers which can be used for its early detection [2]. The test involves collection of cells lining the transformation zone (where the outer squamous cervical cells meet the inner glandular endocervical cells), preparation of smears from the collected cells, staining them with Papanicolaou stain and analyzing them under microscope to detect any abnormalities. The entire process is time consuming, costly and involves observer biases. So an automated system for screening of Pap smear images will be of particular help for the pathologist. This automated classifier is designed with the following purpose—(i) to reduce observer bias, (ii) to reduce laborious task of pathologist by sieving out the normal cases so that they can concentrate more on suspicious cases, as doctors normally spend lots of time in looking at normal cases specially when camp based screenings are performed, (iii) to quantify the features which may be interpreted by doctors in visual terms which are subjective to the concerned doctor and (iv) effective in terms of cost involved.

A number of commercial systems are available for Pap smear screening like PAPNET, ThinPrep imaging system etc., but application of those advanced techniques are limited because of the cost involved and presence of technical and linguistic gap. Many developing countries still have to rely on conventional screening techniques including India where this study has been carried out. So design and delivery of low cost but highly efficient screening system is on high demand.

Literature related to Pap smear image analysis can be categorized based on four factors—(1) the single cell level or smear level consideration, (2) types of segmentation algorithms adopted for identifying ROI, (3) types of features observed and (4) classification processes analyzed. If one considers the types of level used for analysis, it can be observed that previous researches are generally based on two types of analysis i.e. cell level study [2–5] and smear level study [6–11]. In cell level study each image contains a single cell which is either normal or abnormal cervical cell. If one considers smear level then it includes not only multiple cervical cells but also many unwanted debries which may be present in the background.

Another important characteristic on which literature can be categorized is based on segmentation techniques adopted for identifying ROI. Attempts have been made to segment the ROI based on Water immersion technique [6], Genetic algorithm [12], Fuzzy C Mean [13], Watershed segmentation [14], Joint Optimization technique [7] etc. But all the techniques were somehow incomplete in addressing some of the major issues like debries removal and consideration of adequate numbers of indigenous samples to be tested.

Literature can be also categorized based on different features used to quantify the dysplastic changes in a Pap smear image. Chen et al. [15] extracted 13 features which can describe nucleus size, nucleo cytoplasmic (N/C) ratio, shape and texture. But they have not studied the color features. Genctav et al. [16] performed their work on shape features. Guan et al. [17] and Camargo et al. [18] studied color features using DFT and MPEG 7 descriptors respectively. Using DFT accurate frequency information can be obtained but we loose time information and it has bad convergence. On the other hand in MPEG descriptors employ Discrete Cosine Transforms which has the disadvantage of neglecting correlation from the pixels of neighboring block resulting to undesired blocking artifacts affecting reconstruction of images. The present approach overcomes these drawbacks by using Ripplet Type 1 transform. Moreover consideration of small numbers of features may lead us to select an unwanted class during classification. That is why 121 low level features which can quantify the dysplastic changes were observed for this study.

Classification is the last but most important stage in designing any Decision Support System. Literature reflected that Pap smear classification is of two types—cell classification [13,18,19] and smear classification. Cell classification mainly focuses on single cell level study where researchers aim at classifying an individual cell into Normal Class (includes Normal Superficial (NS), Normal Intermediate (NI) and Normal Columnar (NC) cells), Abnormal Class (includes Light Dysplasia (LD), Moderate Dysplasia (MD), Severe Dysplasia (SD) and Carcinoma-in-situ (CIS)). Smear level classification is difficult as compared to cell level classification as it basically focuses on classifying the whole image/sample to Normal, LSIL (Low grade Squamous Intraepithelial Lesion), HSIL (High grade Squamous Intraepithelial Lesion) and Carcinoma-in-situ (according to the Bethesda system of classification) [20].

1.1. *Challenges and motivations*

A Pap smear sample contains 1000 to 10,000 cells and it is a tedious and time consuming task to manually analyze the cells under a microscope. In a Pap smear there may be different types of cervical cells present like normal superficial cells, normal intermediate cells, normal parabasal cells, normal basal cells and endocervical or normal columnar cells. Along with different types of cervical cells other types of cells called inflammatory cells and Red Blood Cells (RBC) are common which are termed as debris present in digitized image. When cervix is affected by cancer these cervical cells undergo various morphological changes which include changes in terms of shape, color and texture which in turn are termed as “features” by the researchers. But to properly study the changes in cervical cells one has to remove these debris from the image. So some of the challenges that has to be faced during designing such automated model for Pap smear classifications are staining quality, overlapped cytoplasm, presence of debris and poor contrast of some images. We aim to address some of the above mentioned challenges and attempt to design an automated classifier of Pap smear images for early diagnosis of cervical cancer. Another challenge in medical imaging is unavailability of hospital based real indigenous images. Genctav et al. [16] mentioned about the difficulty in collecting indig-