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

Debris removal in Pap-smear images

Patrik Malm[a,*], *Byju N. Balakrishnan*[b], *Vilayil K. Sujathan*[c], *Rajesh Kumar*[b], *Ewert Bengtsson*[a]

[a] *Centre for Image Analysis, Division of Visual Information and Interaction, Department of Information Technology, Uppsala University, Box 337, 751 05 Uppsala, Sweden*

[b] *Centre for Development of Advanced Computing, Thiruvananthapuram, India*

[c] *Regional Cancer Centre, Thiruvananthapuram, India*

ARTICLE INFO

Article history:
Received 31 August 2012
Received in revised form
25 February 2013
Accepted 26 February 2013

Keywords:
Debris removal
Pap-smear
Cervical cancer screening
LBC

ABSTRACT

Since its introduction in the 1940s the Pap-smear test has helped reduce the incidence of cervical cancer dramatically in countries where regular screening is standard. The automation of this procedure is an open problem that has been ongoing for over fifty years without reaching satisfactory results. Existing systems are discouragingly expensive and yet they are only able to make a correct distinction between normal and abnormal samples in a fraction of cases. Therefore, they are limited to acting as support for the cytotechnicians as they perform their manual screening.

The main reason for the current limitations is that the automated systems struggle to overcome the complexity of the cell structures. Samples are covered in artefacts such as blood cells, overlapping and folded cells, and bacteria, that hamper the segmentation processes and generate large number of suspicious objects. The classifiers designed to differentiate between normal cells and pre-cancerous cells produce unpredictable results when classifying artefacts.

In this paper, we propose a sequential classification scheme focused on removing unwanted objects, debris, from an initial segmentation result, intended to be run before the actual normal/abnormal classifier. The method has been evaluated using three separate datasets obtained from cervical samples prepared using both the standard Pap-smear approach as well as the more recent liquid based cytology sample preparation technique. We show success in removing more than 99% of the debris without loosing more than around one percent of the epithelial cells detected by the segmentation process.

1. Introduction

According to the World Health Organization (WHO), cervical cancer is the second most common type of cancer among women, annually killing close to 300,000 world wide. 86% of these deaths occur in developing countries [1]. The main reason behind this discrepancy is the absence in developing countries of organized screening programmes using the Papanicolaou test (Pap test) developed by Dr. Georges Papanicolaou in the 1940s [2].

A *Pap-smear* is prepared by acquiring cellular material from the uterine cervix using a spatula or a brush. The collected material is then smeared on a microscope slide, fixated using a spray fixative and then stained using the *Pap-stain* [3].

* *Corresponding author.*
E-mail address: patrik.malm@cb.uu.se (P. Malm).
URL: http://www.cb.uu.se/~patrik (P. Malm).

0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.02.008

When a smear is analysed under a microscope, trained cytologists can not only find evidence of invasive cancer but also detect certain cancer precursors, allowing for early and effective treatment. If detected early, invasive cancer is curable and the 5-year survival rate is as high as 92% [4].

Although the Pap-smear has shown its worth through decades of use, it is hampered by a number of difficulties, e.g., variable smear thickness, uneven cell distribution, obscuring elements such as blood and inflammatory cells, and variable fixation and staining results. To overcome some of these problems a number of so-called liquid-based cytology (LBC) preparation methods have been developed. Using LBC the sample is immersed in a solution which is then subjected to a number of steps that work to homogenize the sample, remove unwanted components (e.g., red blood cells) and finally deposit a suitable mono-layer sample on a glass slide [5]. The result is a sample that should contain a better representation of the biological material presented in a mono-layer fashion which according to several studies, e.g. Jhala and Eltoum [6], lead to better screening results.

Most screening programmes are based on visual screening performed by cytotechnicians in specialized laboratories. The screening work is tedious and, often due to fatigue, error prone. Because of the hazards of fatigue some recommendations say that a cytotechnician should not work with screening more than 7 h a day and analyse no more than 70 samples per day [7].

To overcome some of the human limitations and hopefully to reduce screening costs several attempts to automate the screening process have been made since the 1950s with varying degree of success. Today there are systems that are able to perform a scan and subsequent analysis of a sample but they all have in common that they require specific sample preparation and are complicated and expensive to run [8].

When analysing a Pap-smear the cytotechnician will look for a number of tell-tale signs that a sample contains evidence of malignancy [3]. Many of these signs are related to the appearance of the cell nuclei (i.e. shape, color, nucleus/cytoplasm ratio, size, chromatin distribution). Because of the importance of nuclear characteristics the main focus when developing automated smear analysis has been on the processes of locating/delineating [9–13] as well as extracting features [13–15] from nuclei. Segmenting nuclei in Pap-smears is then a key process, but it is made very difficult by the same complications that make the smears hard for humans to analyse, i.e., variable smear thickness, obscuring elements, et cetera. The LBC preparation methods will reduce these problems but not remove them altogether.

The early attempts at screening systems used various kinds of automated greyscale thresholding [16,17] but more recent projects have used more complicated approaches. Bergmeir et al. [9] uses Canny edge detection followed by the randomized Hough transform [18]. Bamford and Lovell [10] use a dual active contour algorithm. Lin et al. [11] uses a two group object enhancement technique. Malm and Brun [12] uses Canny edge detection followed by anisotropic curve closing. Gençtav et al. [13] use a form of multi-scale watersheds to generate hierarchical partitioning of nuclei and cytoplasms.

All segmentation algorithms in this context are intrinsically created to avoid picking up unwanted objects, henceforth referred to as debris. Still there will in most cases be many debris objects among the segmented "nuclei". When such debris objects are subject to feature extraction and classification designed to detect signs of malignancy the outcome is more or less random leading to great difficulties in designing a system with sufficiently low false positive and false negative rates. In this paper we propose an initial classification stage with the sole purpose of detecting and removing the debris objects. To the best of our knowledge no previous paper has had that focus.

2. Methods

The objective of the work presented in this paper was to develop a robust method for filtering out debris from an initial segmentation result. The method has been tailored to tackle many of the difficulties present in Pap-smear images (Fig. 1). The approach centres around a sequential elimination scheme (Fig. 2) where objects from an initial segmentation are removed if deemed unlikely to be one of the relevant types of cell nuclei. The benefits of a sequential approach are two-fold. First, it allows for a lower-dimensional decision to be made at each stage, thus reducing the effects of the curse of dimensionality, and second, it makes it possible to place more computationally heavy object descriptors at the end of the pipeline where fewer objects remain.

Where applicable, a standard Bayesian quadratic classifier [19] has been used. Furthermore, since each step of the method only tackles a limited number of features, the complexity of the classifier is not as critical.

In the initial step of the proposed method, objects are thresholded based on their area (see Section 2.1). Following the basic thresholding, objects are analysed based on their shape. The second step of the algorithm evaluates objects using region-based and contour-based shape representations (see Section 2.2). The third step constitutes a custom algorithm that measures the elliptical deviation (see Section 2.3). Remaining objects at this stage are evaluated based on their texture (see Section 2.4) and finally their average greyvalue (see Section 2.5).

2.1. *Area*

Area is perhaps the most basic feature available and also the first one used within the field of automated cytology to separate cells from debris [20]. This is of course not an inherently specific feature but for segmentation algorithms where size is not taken into account implicitly, e.g. [12], it is a necessary one.

Finding a lower size threshold is generally not an issue in automated cervical cytology applications. Because it is such a well studied field much prior knowledge regarding cell characteristics, such as average size distribution, is available [21]. However, one of the key changes a cancerous cell undergoes is the substantial increase of nuclear size [3] (Fig. 3). Therefore, determining an upper size threshold that does not systematically exclude diagnostic cells is much harder.

The method described in this paper only uses a lower size threshold to avoid the exclusion of diagnostic cells. The