Pattern Recognition 45 (2012) 4151–4168

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Unsupervised segmentation and classification of cervical cell images

Aslı Gençtav [a,1], Selim Aksoy [a,*], Sevgen Önder [b]

[a] Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey
[b] Department of Pathology, Hacettepe University, Ankara 06100, Turkey

ARTICLE INFO

Article history:
Received 1 August 2011
Received in revised form
18 April 2012
Accepted 2 May 2012
Available online 18 May 2012

Keywords:
Pap smear test
Cell grading
Automatic thresholding
Hierarchical segmentation
Multi-scale segmentation
Hierarchical clustering
Ranking
Optimal leaf ordering

ABSTRACT

The Pap smear test is a manual screening procedure that is used to detect precancerous changes in cervical cells based on color and shape properties of their nuclei and cytoplasms. Automating this procedure is still an open problem due to the complexities of cell structures. In this paper, we propose an unsupervised approach for the segmentation and classification of cervical cells. The segmentation process involves automatic thresholding to separate the cell regions from the background, a multi-scale hierarchical segmentation algorithm to partition these regions based on homogeneity and circularity, and a binary classifier to finalize the separation of nuclei from cytoplasm within the cell regions. Classification is posed as a grouping problem by ranking the cells based on their feature characteristics modeling abnormality degrees. The proposed procedure constructs a tree using hierarchical clustering, and then arranges the cells in a linear order by using an optimal leaf ordering algorithm that maximizes the similarity of adjacent leaves without any requirement for training examples or parameter adjustment. Performance evaluation using two data sets show the effectiveness of the proposed approach in images having inconsistent staining, poor contrast, and overlapping cells.

1. Introduction

Cervical cancer is the second most common type of cancer among women with more than 250,000 deaths every year [1]. Fortunately, cervical cancer can be cured when early cancerous changes or precursor lesions caused by the Human Papilloma Virus (HPV) are detected. However, the cure rate is closely related to the stage of the disease at diagnosis time, with a very high probability of fatality if it is left untreated. Therefore, timely identification of the positive cases is crucial.

Since the discovery of a screening test, namely the Pap test, introduced by Dr. Georges Papanicolaou in the 1940s, a substantial decrease in the rate of cervical cancer and the related mortality was observed. The Pap test has been the most effective cancer screening test ever, and still remains the crucial modality in detecting the precursor lesions for cervical cancer. The test is based on obtaining cells from the uterine cervix, and smearing them onto glass slides for microscopic examination to detect HPV effects. The slides are stained using the Papanicolaou method where different components of the cells show different colors so that their examination becomes easier (see Figs. 1 and 2 for examples).

There are certain factors associated with the sensitivity of the Pap test, and thus, the reliability of the diagnosis. The sensitivity of the test is hampered mostly by the quality of sampling (e.g., number of cells) and smearing (e.g., presence of obscuring elements such as blood, mucus, and inflammatory cells, or poorly fixation of specimens). Both intra- and inter-observer variabilities during the interpretation of the abnormal smears also contribute to the wide variation in false-negative results [2]. The promise of early diagnosis as well as the associated difficulties in the manual screening process have made the development of automated or semi-automated systems that analyze images acquired by using a digital camera connected to the microscope an important research problem where more robust, consistent, and quantifiable examination of the smears is expected to increase the reliability of the diagnosis [3,4].

Both automated and semi-automated screening procedures involve two main tasks: segmentation and classification. Segmentation mainly focuses on separation of the cells from the background as well as separation of the nuclei from the cytoplasm within the cell regions. Automatic thresholding, morphological operations, and active contour models appear to be the most popular and common choices for the segmentation task in the literature. For example, Bamford and Lovell [5] segmented the nucleus in a Pap smear image using an active contour model that was estimated by using dynamic programming to find the boundary with the minimum cost within a bounded space around the darkest point in the image. Wu et al. [6] found the boundary

* Corresponding author. Tel.: +90 312 2903405; fax: +90 312 2664047.
E-mail addresses: asli@ceng.metu.edu.tr (A. Gençtav), saksoy@cs.bilkent.edu.tr (S. Aksoy), sonder@hacettepe.edu.tr (S. Önder).
[1] Present address: Department of Computer Engineering, Middle East Technical University, Ankara 06531, Turkey.

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.05.006

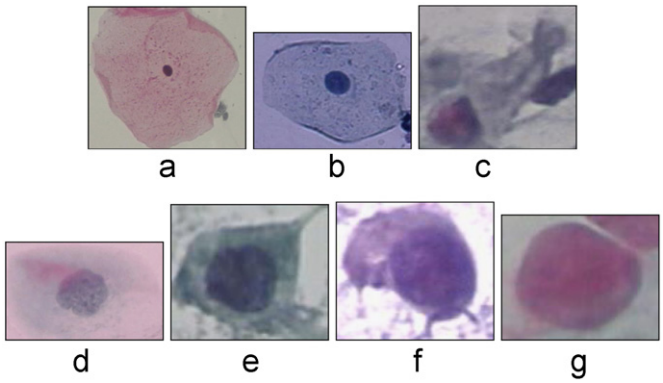

Fig. 1. Examples from the Herlev data set involving a single cell per image. The cells belong to (a) superficial squamous, (b) intermediate squamous, (c) columnar, (d) mild dysplasia, (e) moderate dysplasia, (f) severe dysplasia, and (g) carcinoma in situ classes. The classes in the first row are considered to be normal and the ones in the second row are considered to be abnormal. Average image size is 156×140 pixels. Details of this data set are given in Section 2.

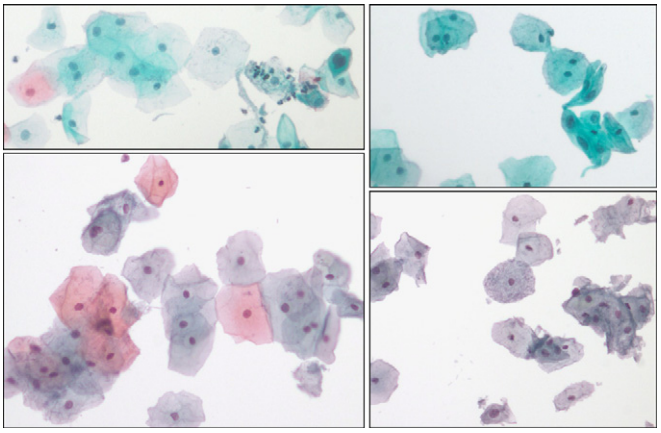

Fig. 2. Examples from the Hacettepe data set involving multiple overlapping cells with inconsistent staining and poor contrast that correspond to a more realistic and challenging setting. Details of this data set are given in Section 2.

of an isolated nucleus in a cervical cell image by using a parametric cost function with an elliptical shape assumption for the region of interest. Yang-Mao et al. [7] applied automatic thresholding to the image gradient to identify the edge pixels corresponding to nucleus and cytoplasm boundaries in cervical cell images. Tsai et al. [8] replaced the thresholding step with *k*-means clustering into two partitions. Dagher and Tom [9] combined the watershed segmentation algorithm with the active contour model by using the watershed segmentation result of a down-sampled image as the initial contour of the snake for the segmentation of blood cells and corneal cells. Huang and Lai [10] also used the marker-based watershed segmentation algorithm to find an approximate segmentation, applied heuristic rules to eliminate non-nuclei regions, and used active contours to improve the nuclei boundaries in biopsy images of liver cells. Harandi et al. [11] used the active contour algorithm to identify the cervical cell boundaries, applied thresholding to identify the nucleus within each cell, and then used a separate active contour for each nucleus to identify the corresponding cytoplasm within connected cell groups. Plissiti et al. [12] detected the locations of nuclei centroids in Pap smear images by using the local minima of image gradient, eliminated the candidate centroids that were too close to each other, and used a support vector machine (SVM) classifier for the final selection of points using color values in square neighborhoods. Then, they used the detected centroids as markers in marker-based watershed segmentation to find the nuclei boundaries, and eliminated the false-positive regions by using a binary SVM classifier with shape, texture, and intensity features [13]. Li et al. [14] used *k*-means clustering for a rough partitioning of an image into nucleus, cytoplasm, and background areas, and then performed snake-based refinement of nucleus and cytoplasm boundaries.

Most of these methods focus on the segmentation of only the nuclei [5,6,9,10,12,13] for which there is relatively higher contrast around the boundaries. However, detection of the cytoplasm regions is also crucial because cytoplasm features have been shown to be very useful for the identification of abnormal cells [15]. Even so, these nuclei-specific methods do not necessarily generalize well for the detection of cytoplasms that create an increased difficulty due to additional gradient content and local variations. Furthermore, many of the methods [5–8,14] assume a single cell in the input image where there is only one boundary (nucleus) or at most two boundaries (nucleus and cytoplasm) to detect as in the examples in Fig. 1. However, this is usually not a realistic setting as can be seen in the images in Fig. 2 where one cannot make any assumption about the number of cells or expect that these cells appear isolated from each other so that they can be analyzed independently. Among the proposed methods, automatic thresholding for nuclei detection [7,8] assumes a bimodal distribution but can only be used for the segmentation of isolated cells. Watershed-based methods [10,12,13] can identify more details and can handle multiple cells but have the potential of over-segmentation, so they require carefully adjusted preprocessing steps or carefully selected markers. Active contour-based methods [5,9–11,14] can provide a better localization of boundaries when there is sufficient contrast but are often very difficult to initialize with a very sensitive process for parameter and capture range selections. Our earlier work showed that it can be very difficult to find reliable and robust markers for marker-based watershed segmentation and very difficult to find a consistent set of parameters for active contour models when there are multiple cells in the image [16]. A recent development of interest has been the work on the incorporation of shape priors into active contour models to resolve overlapping and occluded objects [17]. However, these priors have been mainly applied to the segmentation of objects with a well-defined and consistent appearance, whereas it is not straightforward to define a shape prior for the overlapping cells with highly varying cytoplasm areas as shown in Fig. 2. Moreover, their initialization is still an existing problem when the number of cells and their approximate locations are unknown.

In this paper, our first major contribution is a generic and parameter-free segmentation algorithm that can delineate cells and their nuclei in images having inconsistent staining, poor contrast, and overlapping cells. The first step in the segmentation process separates the cell regions from the background using morphological operations and automatic thresholding that can handle varying staining and illumination levels. Then, the second step builds a hierarchical segmentation tree by using a multi-scale watershed segmentation procedure, and automatically selects the regions that maximize a joint measure of homogeneity and circularity with the goal of identifying the nuclei at different scales. The third step finalizes the separation of nuclei from cytoplasm within the segmented cell regions by using a binary classifier. The proposed algorithms are different from related work in that (1) the automatic thresholding step can handle multiple cell groups in images because the gray scale bimodality assumption holds when the goal is to extract only the background, and (2) no initialization, parameter adjustment, or marker detection are required. Unlike some other approaches that tried to select a single scale from watershed hierarchies [18] or use thresholds on region features to select a subset of regions