Article can be accessed online at http://www.publishingindia.com

Probabilistic Segmentation Methods for Early Detection of Uterine Cervical Cancer

Abhishek Das *, Avijit Kar, Debasis Bhattacharyya*****

Abstract

Uterine Cervical Cancer is one of the prevalent forms of cancer in women worldwide. Most cases of cervical cancer can be prevented through screening programs aimed at detecting precancerous lesions. In this paper, novel methods have been proposed for automated probabilistic image segmentation of cervical cancer. The detection of cervical lesions is an important issue in image processing because it has a direct impact on surgical planning. We examined the segmentation accuracy based on a validation metric against the estimated composite latent gold standard, which was derived from several experts' manual segmentations. The distribution functions of the lesion and control pixel data were parametrically assumed to be a mixture of probability distributions with different shape parameters. We also estimated the corresponding receiver operating characteristic (ROC) curve over all possible decision thresholds. The automated segmentation yielded satisfactory accuracy with protean optimal thresholds.

Keywords: Segmentation, Clustering, Gaussian Mixture Model

1. Introduction

Cervical Cancer is one of the prevalent forms of cancer afflicting female population in developing countries worldwide. It is, however, now ranked 11th in incidence and 13th in mortality [JCan05] in the developed countries, due to the ability to detect the precancerous lesions through government-sponsored Cervical Cancer Screening Program and to treat them in time. This success in the developed countries has been achieved through a synergy between Cervical Cytology Screening for Cervical Intra-epithelial Neoplasia (CIN) before they become invasive and their effective treatment directed by colposcopy.

Colposcopy is a very effective, non-invasive diagnostic tool. In colposcopy, the cervix is examined non-invasively by a colposcope which is a specially designed binocular stereo-microscope. The abnormal cervical regions turn to be white after application of 5% acetic acid and are called Acetowhite (AW) lesions which are then biopsied under colposcopic guidance for confirmation by histo-pathological examination. Modern colposcopes can produce a digital image of the cervix. Colposcopy today is considered the *gold standard* for detection and treatment of pre-cancerous lesions of the cervix [JCan05]. The purpose of this study is to explore whether digital colposcopy, combined with recent advances in camera technology and automated image processing, could provide an inexpensive alternative to Pap screening and conventional colposcopy. The goal of this work is to develop automated probabilistic cervical lesion segmentations. Our methods are developed mainly under the pixel independence assumptions.

2. Automated System

An automated system is proposed in this paper that is used for diagnosis of CIN. The scheme is presented as a block diagram in Fig. 1. The process of translating raw cervix image acquired using a Digital Colposcope into a thorough diagnosis of CIN is decomposed into four modules: 1) removal of Specular Reflection (SR) from raw

* Department of Information Technology, Tripura University (A Central University), Agartala, Tripura, India.
E-mail: adas@tripurauniv.in
** Department of Computer Science & Engineering, Jadavpur University, Kolkata, West Bengal, India.
*** Department of Gynecology & Obstetrics, College of Medicine & SDM Hospital, Kolkata, West Bengal, India.

cervigrams; 2) segmentation of cervix region of interest (ROI); 3) segmentation of cervix ROI into acetowhite (AW), columnar epithelium (CE) and squamous epithelium (SE); 4) classification of AW regions into AW, mosaic, or punctuation tiles;

3. Algorithms for Segmentation

Salient features observed in cervical images consist of the cervix ROI, SR, AW, SE, and CE. While the AW region is the single most important region for detecting the presence and extent of CIN, the ROI must also be extracted[IST11]. SR, which adversely affects the AW segmentation process, must be removed[BMV11]. During the AW segmentation procedure, the CE and SE are designated as distinct regions. The classification process of these macro features constitutes the first three modules of the automated diagnosis system. The Algorithms for pre-processing[IV11] has been elaborated in our previous work.

In our problem to segment the Region of Interest (ROI) for the CE, SE & AW regions, the underlying probability density function has to be estimated from the available data. There are various ways to approach the problem. Sometimes we may know the type of the pdf (eg. Gaussian, Rayleigh) but we do not know certain parameters, such as the mean value or the variance. In other cases we may not have information about the type of the pdf but we know certain statistical parameters, such as the mean value and the variance. Depending on the available information, different approaches can be adopted. Considering the cervigram as a Gaussian Mixture Model (GMM), the probabilistic segmentation algorithm proposed is a variant of the Expectation-Maximization (EM) algorithm.

To estimate an unknown probability density function $p(x)$ is via a linear combination of probability density functions[TSA12] in the form of

$$p(x) = \sum_{j=1}^{J} p(x \mid j)\, P_j$$

where as

$$\sum_{j=1}^{J} P_j = 1. \qquad \int_x p(x \mid j)\, dx = 1.$$

It can be assumed that there are J distributions contributing to the formation of $p(x)$. Thus this mathematical modeling assumes that each point x may be drawn from any of the J model distributions with probability P_j , $j = 1,2,\ldots J$. The first step of the procedure involves the choice of the set of density components $p(x \mid j)$ in parametric form, that is, $p(x \mid j;\ \theta)$, and then the computation of the unknown parameters, θ and P_j , $j = 1,2,\ldots J$. based on the set of the available training samples x_k . There are various ways to achieve this.

By maximizing the likelihood function $\prod_{\mathbf{k}} p(x_k\ ;\ \theta,\ P_1, P_2,\ldots\ldots,P_j)$ with respect to θ and the Pj 's can be thought first. The problem arises from the fact that the parameters which are not known enter the maximization task in a *nonlinear fashion*; thus, nonlinear optimization iterative techniques have to be adopted. We will focus on the EM algorithm to overcome this difficulty.

The Expectation-Maximization (EM) algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. In ML estimation, we estimate the model parameter(s) for which the observed data are the most likely. Each iteration step of the EM algorithm consists of two processes: The E-step, and the M-step. In the expectation, or E-step, the missing data are estimated given the observed data and current estimate of the model parameters. This is achieved using the conditional expectation.

In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step is used in lieu of the actual missing data. Convergence is assured since the algorithm is guaranteed to increase the likelihood [TSA12], [PR09] at each iteration.

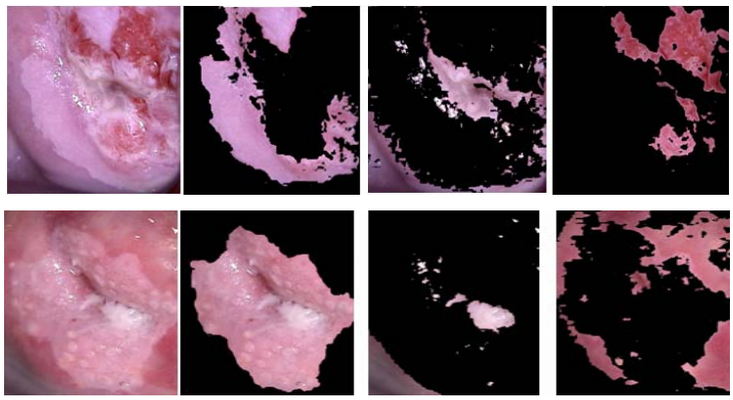
Let us first hypothesize the problem in more general terms and then apply it to our specific task. Let us denote by y *the complete data samples*, with $y \in Y \subseteq R^m$, and let the corresponding pdf be $py\ (y;\ \theta)$, where θ is an unknown parameter vector. The samples $\boldsymbol{y}$ however cannot be directly observed. What we observe instead are samples $x = g(y) \in X_{ob} \subseteq R^l$, $l < m$. We denote the corresponding probability density function $px\ (x;\ \theta)$. This is usually n to 1 mapping.

Let $Y(x) \subseteq Y$ be the subset of all the y's corresponding to a specific x. Then the probability density function of the incomplete data is given by

$$px\ (x;\ \theta) = \int_{Y(x)} py\ (y;\ \theta)\, dy$$

Figure 1: Cervigram, Segmented AW Region, Segmented CE Region, Segmented SE region

(left to right)



As we already know, the maximum likelihood estimate of θ is given by

$$\hat{\theta}_{ML}: \sum_k = 0$$

However the y's are still not available. Hence the variant of the Expectation-Maximization algorithm maximizes the expectation of the log likelihood function, conditioned on the observed samples and the current iteration estimate of θ. The two steps of the modified algorithm are:

- E-step: At the (t+1)th step of the iteration, where θ(t) is available, compute the *expected value of*

$$Q(\theta; \theta(t)) \equiv E\ [\ \sum_k \ln (py\ (y_k; \theta \mid X: \theta(t))\ )\ ]$$

This is the so called expectation step of the algorithm.

- M-step: Compute the next $(t+1)^{th}$ estimate of θ by *maximizing Q(θ: θ(t)),*

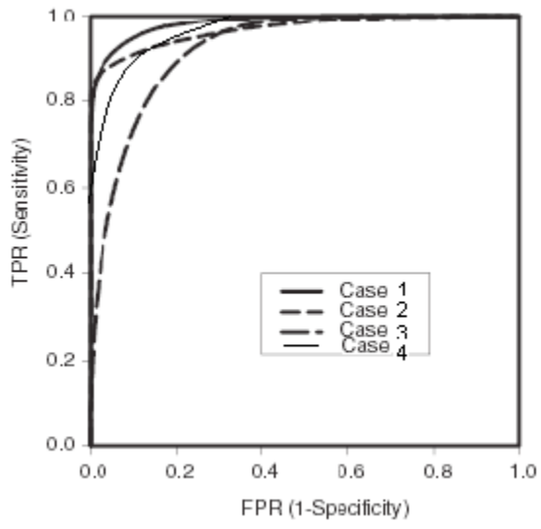
that is, $\theta(t+1) : = 0$

This is the maximization step, where differentiability has been assumed as obvious. To apply the EM algorithm, we start from an initial estimate θ(0) and iterations are terminated if $|\theta(t+1) - \theta(t)| \leq \epsilon$ for an aptly chosen vector norm and ϵ.

4. Results

We assume the ROI Cervigram to be a mixture of Gaussian density function, and using the modified EM algorithm, we get the segmentation of the Aceto White lesions as shown in Fig 1. The Receiver Operating Characteristic

Figure 2: Estimated ROC Curves of the Four Cervical Lesion Cases



curve (ROC) of 4 cancer patients out of 750 samples to detect the cancerous lesion regions have been presented in Fig 2. A ROC curve shows true positive rate versus false positive rate (equivalently, sensitivity versus 1–specificity) for different thresholds of the classifier output. We can use it, for example, to find the threshold that maximizes the classification accuracy or to assess, in more broad terms, how the classifier performs in the regions of high sensitivity and high specificity.

The accuracy of a diagnostic test can be summarized in terms of an ROC curve. It is a plot of sensitivity (true lesion fraction) vs (1-specificity) (true non-lesion fraction). There is always a trade-off between these two false positive and false negative error rates, or specificity and sensitivity.

5. Summary

In this work, we have presented systematic approaches to automate cervical images leading to pixel-wise probabilistic segmentation of the lesion class. We developed a protean of the EM algorithm for estimating the latent gold standard. In addition, we modeled the probabilistic segmentation results using a mixture of distributions with different shape parameters. Summary accuracy measures including ROC curve were also estimated to validate our segmentation results.

References

1. Parkin, D. M., Bray, F., Ferlay J. & Pisani P. (2005). Global cancer statistics.*CA: A Cancer Journal for Clinicians,* 55(2), 74-108. [JCan05]
2. Das, A., Kar, A. & Bhattacharyya, D. (2011). Elimination of specular reflection and identification of ROI: The first step in automated detection of uterine cervical cancer using digital Colposcopy. *IEEE Imaging Systems & Techniques* ISSN 1550-6037 pp 237-141, 2011 [IST11]
3. Das, A., Kar, A. & Bhattacharyya, D. (2011). *Preprocessing for Automatic Detection of Cervical Cancer.* Paper presented at 15th International Conference on Information Visualization, University of London, U.K. [IV11]
4. Das, A., Kar, A. & Bhattacharyya, D. (2011). Biomedical visualization. In Banissi, E. et al (Eds.), *In Information Visualization* (pp. 597-600). USA: IEEE Computer Society. [BMV11]
5. Das, A., Kar, A. & Bhattacharyya, D. (2012). *Implication of Technology on Society in Asia: Automated Detection of Cervical Cancer.* Paper presented at IEEE Conference on Technology and Society in Asia (T&SA), Singapore. [TSA12]
6. *Theodoridis, S. & Koutroumbas, K. (2009). Pattern Recognition*. New York: Elsevier. [PR09]