A Probabilistic Approach to Segmentation and Classification of Neoplasia in Uterine Cervix Images Using Color and Geometric Features

Yeshwanth Srinivasan[a], Dana Hernes[a], Bhakti Tulpule[a], Shuyu Yang[a], Jiangling Guo[a] , Sunanda Mitra[a], Sriraja Yagneswaran[a], Brian Nutter[a], Jose Jeronimo[b] Benny Phillips[c], Rodney Long[d], Daron Ferris[e]

[a]Department of Electrical and Computer Engineering, Texas Tech University, Lubbock TX 79409;
[b]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20852;
[c]Lubbock Gynecologic Oncology Associates, Lubbock, TX 79410;
[d]National Library of Medicine, Rockville, MD 20852;
[e]Medical College of Georgia, Augusta, Georgia 30912

ABSTRACT

Automated segmentation and classification of diagnostic markers in medical imagery are challenging tasks. Numerous algorithms for segmentation and classification based on statistical approaches of varying complexity are found in the literature. However, the design of an efficient and automated algorithm for precise classification of desired diagnostic markers is extremely image-specific. The National Library of Medicine (NLM), in collaboration with the National Cancer Institute (NCI), is creating an archive of 60,000 digitized color images of the uterine cervix. NLM is developing tools for the analysis and dissemination of these images over the Web for the study of visual features correlated with precancerous neoplasia and cancer. To enable indexing of images of the cervix, it is essential to develop algorithms for the segmentation of regions of interest, such as acetowhitened regions, and automatic identification and classification of regions exhibiting mosaicism and punctation. Success of such algorithms depends, primarily, on the selection of relevant features representing the region of interest. We present color and geometric features based statistical classification and segmentation algorithms yielding excellent identification of the regions of interest. The distinct classification of the mosaic regions from the non-mosaic ones has been obtained by clustering multiple geometric and color features of the segmented sections using various morphological and statistical approaches. Such automated classification methodologies will facilitate content-based image retrieval from the digital archive of uterine cervix and have the potential of developing an image based screening tool for cervical cancer.

Keywords: Cervical cancer, automatic segmentation, uterine cervix, image morphology

1. INTRODUCTION

Cervical cancer is one of the most common forms of cancer afflicting women, affecting about 12,200 women in America [1] and 380,000 women worldwide [2] every year. However, it can be cured if it is detected during its early stages and treated appropriately [3]. The traditional methods of detecting cervical cancer is the Pap smear test [4] in which exfoliated cells from the cervix are scraped out and observed under a microscope for squamous and glandular intraepithelial lesions (SIL). This technique, apart from being not very accurate, requires trained personnel to evaluate the samples. Also, the Pap smear is only a screening test and not a diagnostic tool [5].

Medical Imaging 2005: Image Processing, edited by J. Michael Fitzpatrick,
Joseph M. Reinhardt, Proc. of SPIE Vol. 5747 (SPIE, Bellingham, WA, 2005)
1605-7422/05/$15 · doi: 10.1117/12.597075

Most cases of mortality in cervical cancer occur in remote areas of developing countries, where trained personnel are not available for conducting appropriate screening, diagnostic, and follow-up tests. This calls for inexpensive and automatic methods of screening tools for initial evaluation [6-12]. Digital colposcopy [13-16] lends itself extremely well to this requirement. It is a non-invasive diagnostic tool in which the cervix is examined under a colposcope, which is a specially designed binocular microscope, after the application of acetic acid. The abnormal region appears whiter than the normal region, and is called the acetowhite (AW) region. Abnormal blood vessel changes can also be observed under the colposcope. Modern colposcopes also produce a digital image of the cervix. However, due to the absence of appropriate image processing software for processing the images acquired in commercial colposcopes, the colposcopic image still needs to be evaluated by a trained physician, who then suggests a biopsy to confirm the presence of pre-cancer or cancer.

In recent years, several attempts have been made to study the colposcopic image, extract features and automatically detect the presence of precancerous lesions [13-15]. Building a full-fledged automatic diagnostic tool using appropriate pattern recognition and classification algorithms is complicated by several factors such as non-uniform illumination, lack of good quality, high resolution images, glare effects, lack of repeatability, non-uniform features and others, and to our knowledge design of such a system has not yet been achieved. The National Library of Medicine and the National Cancer Institute (NCI) are creating an archive of 60,000 digitized images of the uterine cervix from cervicographic film images. These images were collected from the NCI Guanacaste Project for the study of visual features correlated with the development of precancerous neoplasia in a population that has been identified as having a high rate of cervical cancer. These images in their uncompressed form have a file size of 16 MB each. NLM is developing tools for the analysis and dissemination of these images and the collected clinical data over the Web.

In this transaction, we present our work on developing a fully automatic system to detect the presence of mosaic patterns formed by cervical intraepithelial neoplasia (CIN) and segment such regions. It is unique in the sense that it includes algorithms to segment the acetowhite region, and check segments of the acetowhite region for presence of mosaic patterns using morphological filters and a simple, yet effective, geometrical and color features-based clustering to identify the regions affected by mosaicism. An evaluation of the pathology can be made based on the results of the clustering scheme.

The rest of this paper is organized as follows. Section 2 explains the importance of the acetowhite regions and the method used to segment it from the rest of the cervix image. Section 3 provides a brief review of the morphological ROSE (rotating structuring element) filters [17, 18] and its significance in extracting the geometrical features. Section 4 explains the need for using color features. Section 5 presents the results of clustering using various combinations of geometrical and color features and an evaluation of these results. Section 6 concludes the paper.

2. ACETOWHITE SEGMENTATION

The acetowhite region is the most important of all colposcopic observations. Application of acetic acid to cervix surfaces causes dehydration of the abnormal cervical cells, whose big nucleus make difficult for the light to pass through the cells, thus making all forms of Cervical Intraepithelial Neoplasia (CIN) show some degree of opacity or acetowhiteness [16]. The various patterns observed with abnormal epithelium, like vasculature, mosaic and punctations, are observed inside this acetowhite region. It is therefore essential to segment the acetowhite region from the entire colposcopic image first before looking for these patterns. It also serves to greatly reduce the search space for the patterns as the important acetowhite region is only a fraction of the entire image of a cervix.

Acetowhite regions having degrees of opaqueness show pixel brightness and texture that differentiate them from surrounding normal tissues. Therefore, pixel gray level can be used as a feature to separate the acetowhite region by applying appropriate pattern recognition technique. Because of the existence of non-uniform illumination during the image acquisition process and the uneven surface of the cervix, pixel-based classification is sensitive to brightness reduction in areas away from the cervix opening, which is usually the center of the image.

First, the region containing the cervix area is cropped from the entire image. Then image enhancement techniques are applied before classification of pixels in order to reduce the effect of non-uniform illumination. An advanced pattern recognition technique, namely, deterministic annealing [19] is used for pixel classification. The details of this procedure can be found in reference [15, 19]. After classification, the acetowhite regions are extracted by identifying pixels belonging to the same group.

3. FEATURE EXTRACTION

Vascular patterns associated with abnormal epithelium are observed inside the acetowhite region. These patterns include mosaic, punctations and atypical vessels generally referred to as vasculature. In this transaction we restrict ourselves to identifying only the mosaic patterns. The mosaic patterns are due to capillaries looping around the surface tissues and appearing to be in parallel to the surface. The mosaic vessels are usually a regular pattern of intertwining strands of dilated capillaries, and are readily observed as a distinct, dark red and regular pattern in the acetowhite region.

The method used to determine the mosaic cell structure in the cervical lesion image is to use a morphological opening operator with the linear rotating structuring element, ROSE [17]. Since the succeeding operations apply only to gray scale images, the first step is to convert the RGB color image to an intensity image with values in the range [0, 255]. The cell structures in the image are darker in color (values closer to zero) than the surrounding background. Since the morphological operations assume the cell structures, which are the objects of interest, to have a lighter color than the background, the image is complemented before morphological processing. The top-hat transform is used to segment objects in gray scale images, and in this case is used to separate the cell structure from the background. This transform is especially useful for images containing varying amounts of gray scale.

The cell structures are oriented in multiple directions in the image making it necessary to use more than one structuring element. Eight different structuring elements at eight different angles, 22.5 degrees apart, are used to determine the cell structures. These constitute the linear rotating structuring element (ROSE). A sequence of opening operations is performed on the top hat transformed image using each of these structuring elements, and the resulting images are combined. To determine the underlying vessel structure, this image is thresholded using the Otsu method [20] to yield a binary image and then skeletonized to yield a final image in which the white pixels correspond to underlying vessel structure. More details on ROSE can be found in [17,18].

Before extracting the geometrical features, an additional spur operation is performed to remove the spur pixels. This helps to clean the cells off spurious edges and the cell area, height and width can be calculated accurately. Since the number of black pixels enclosed in each bounded cell is directly proportional to the actual area of the cell in the vessel structure, we use this count as a measure of the area of the cells. Examples of cell structures obtained by application of ROSE are shown in section 5.

4. COLOR FEATURES

Geometrical features like cell area, height and width are direct measures of the intercapillary distance, which in turn is an indicator of the severity of the lesion. These features are expected to be significantly different for normal cervix tissue and lesions containing mosaic patterns. The cell area, for example, should be much larger for lesions containing mosaic patterns than for normal cervix tissue. However, due to non-uniformities in the imaging process, these values vary over a large range. The non-uniformities include camera position, excess illumination, stretching or contraction of lesions during imaging, etc. These non-uniformities weigh down the effectiveness of the geometrical features.

Physicians observe the color changes in the vessels to make conclusions about the severity of the lesions. It is therefore intuitive to use the color features to automatically classify regions into mosaic and non-mosaic regions. The YCbCr color space is used to measure the luminance and chrominance values of a pixel [21]. The Y value indicates the luminance and the blue and red chrominance values are measured by Cb and Cr, respectively. The Cb and Cr values for the cervical lesion image remain relatively uniform throughout. The luminance value, Y, vary significantly which gives the appearance of the color changes.

Initial visual analysis of a cervical lesion image shows that the sections of the image containing the mosaic features have darker pixel values than the surrounding features in the image. Therefore in the YCbCr color space, the Cb and Cr values for mosaic sections will at least be minimally different from that for non-mosaic sections. Y, Cb and Cr values can be directly obtained from the RGB values using the following equations:

$$Y = 0.299R + 0.587G + 0.114 B \tag{1}$$
$$Cb = -0.169R - 0.331G + 0.500B \tag{2}$$
$$Cr = 0.500R – 0.419G – 0.081B \tag{3}$$

The mosaic and non-mosaic regions also show variations in their RGB values. In the RGB color space, the green and blue values are higher for the non-mosaic points than for the points in the sections containing the mosaic features in the image. This is because mosaic structure corresponds to cutaneous vessels which are heavily saturated with red, and have much smaller green and blue values compared to the sub-cutaneous vessels observed in other acetowhite regions.

For our application, we used two different sets of color features – Cb and Cr from the YCbCr space and G and B values from the RGB space. The results for classification using these features are presented in the subsequent section.

5. RESULTS

In this section we present step-by-step results to illustrate the evolution of our technique. The overall procedure can be summarized as follows. The colposcopic or the cervicographic image is first treated with the methods described in section 2 to segment the acetowhite region. The acetowhite region is then divided into smaller regions. Since the original images were very large, this region was divided into 25 smaller sections. Each of these sections is then subjected to geometrical feature extraction procedure using ROSE, as outlined in section 3. The average area of all the cells in each section is the only geometrical feature used. For all the white points in the binary image obtained after applying ROSE, the corresponding green, blue, Cb and Cr values are determined from the original image. The average values of each of these components are recorded for each section along with the average area of the cells.

Clustering is used to group the sections of the acetowhite image into two groups: mosaic and non-mosaic. The k-means clustering [22] technique was chosen for its simplicity and speed of operation. Two sets of three dimensional vectors were formed from the extracted features for each section. The first set includes the average area of the cells, average Cb values and the average Cr values for all the pixels determined to be a part of the vessel structure. The second set includes the average area of the cells, average green values and average blue values for all the pixels determined to be a part of the vessel structure. The number of groups in this particular case is two. The clustering algorithm examines each component in the image section and assigns it to one of the two groups.

Most of the images used for testing were from a set of 169 images, each of size 2399x1637 pixels, obtained from the National Library of Medicine. Each of these images has been marked for mosaic, punctations and vascular structures by physicians, and hence the ground truth data are available for these images.

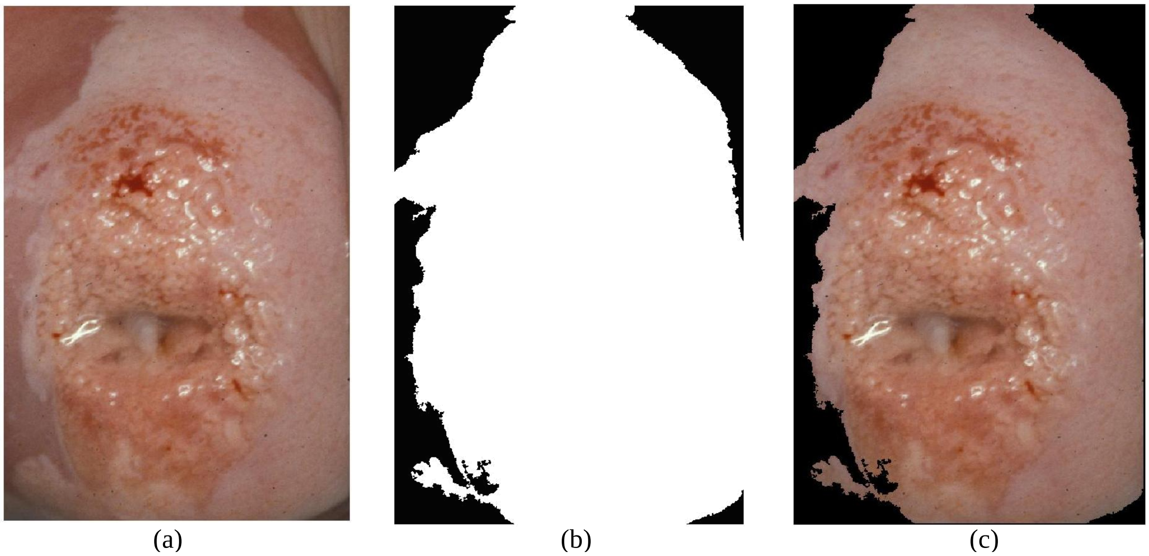


Figure1: (a) Original colposcopic image (b) Binary mask for acetowhite region (c) Segmented acetowhite region.

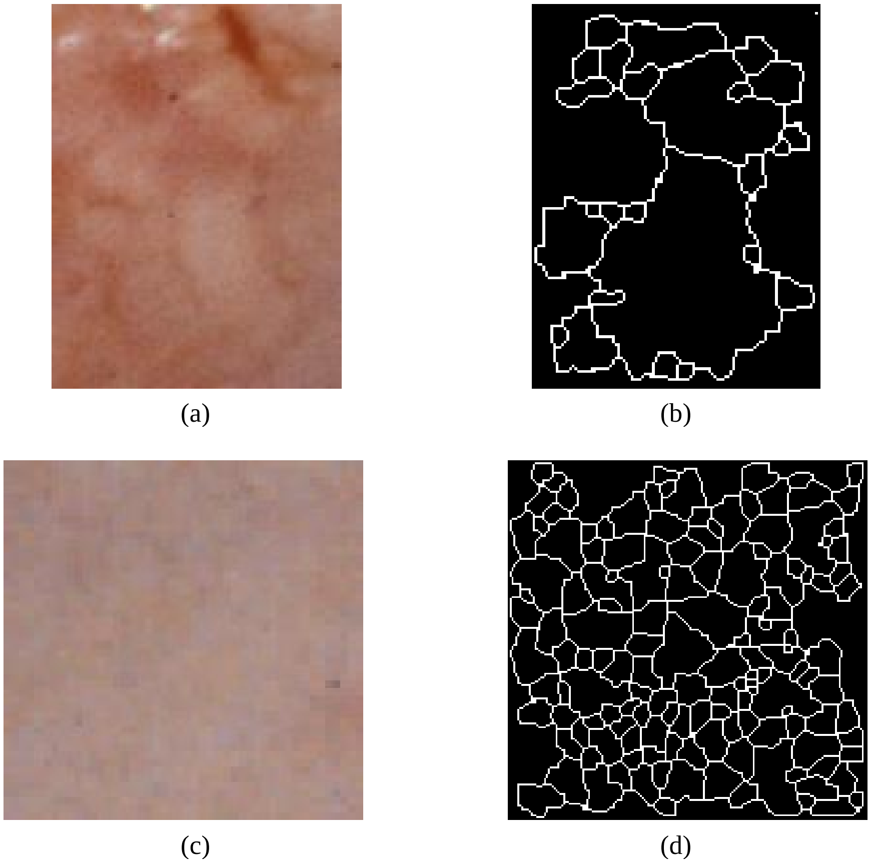


Figure 2: (a) Image section with mosaic pattern (b) 2 (a) after morphological processing (c) Image section with no mosaic pattern (d) 2 (c) after morphological processing.

Figure 1(a)-(c) show the separation of the acetowhite region from the cervix image using deterministic annealing, as explained in sections 2. Since the original image was very large in size, only a 588x892 section of the original image, that includes the important histology, is illustrated. Figure 1(a) shows the original colposcopic image, figure 1(b) shows the binary mask generated by the deterministic annealing scheme and figure 1(c) shows the result of applying the binary mask on the colposcopic image in 1(a) to segment out the acetowhite region.

Figures 2(a)-(d) illustrate the ROSE algorithm. Figure 2(a) shows a section of the image in figure 1(c) that was found to contain mosaic patterns and figure 2(b) shows the same section after morphological operations and thresholding. The mosaic patterns are more clearly visible in this image than in figure 1. Figure 2(c) corresponds to a section of image in figure 1(c) that does not contain any mosaic patterns and figure 2(d) shows the result of applying the procedure in section 3. The acetowhite image was converted to grayscale before the morphological operations were applied.

It is readily seen that the section containing the mosaic pattern (figures 2(a) and (b)) consists of several large cell structures and have significantly greater area than the section containing no mosaic patterns (figures 2(c) and (d)). The average area of the cells in figure 2(b) was found to be 62.875 pixels against that of figure 2(d), which was found to be 54.75 pixels.

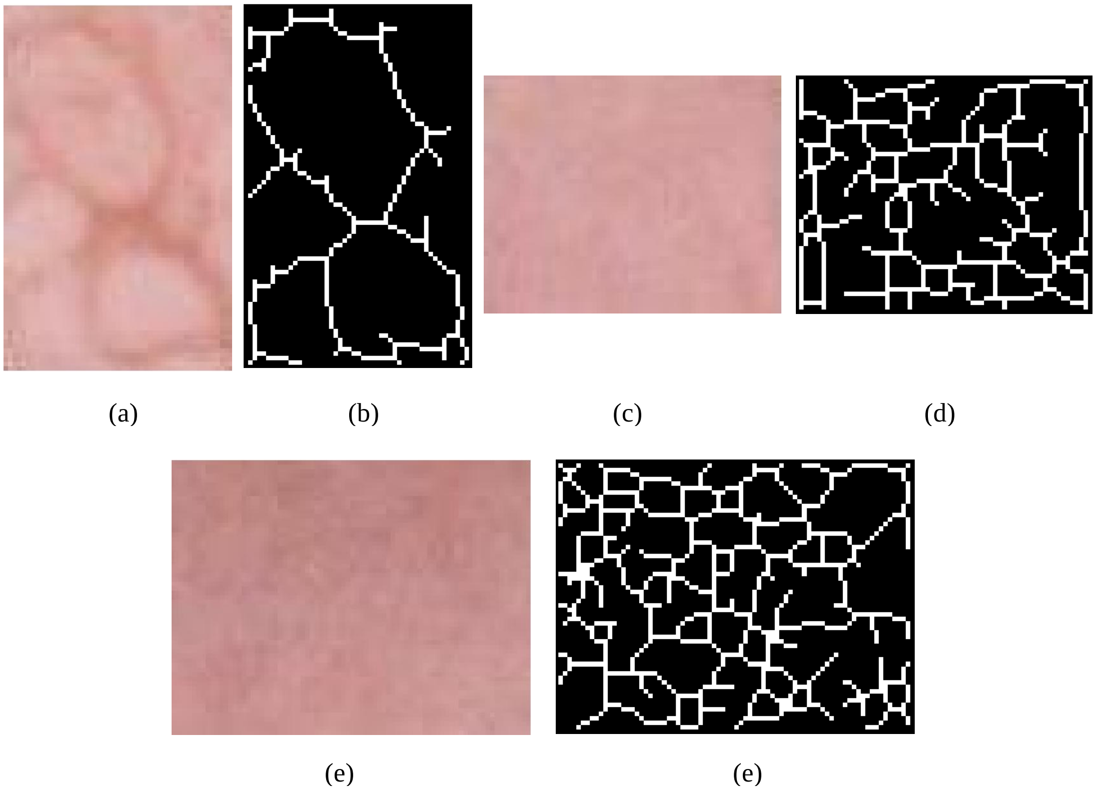


Figure 3: Typical lesions and their extracted vascular structure (a) Mosaic AW region (b) Extracted vascular structure from (a) (c) Non-mosaic AW region (d) Extracted vascular structure from (c) (e) Normal region (f) Extracted vascular structure from (e)

Figure 3 shows some typical regions obtained from cervix images and the result of applying the morphological ROSE operator described. The most interesting results are observed for the normal region, shown in figure 3 (e) and (f), where even the fine vessel structure is extracted. This vessel structure is not immediately visible like the mosaic pattern, and since it lies outside the AW region it is not considered precancerous. Although it needs to be verified, we believe, that it is a pointer towards non-invasive diagnosis of cervical cancer.

The Cb, Cr, green and blue values were extracted for the white pixels in the morphologically filtered image sections, like the ones shown in figure 2(b) and (d). The average area, Cb and Cr values of four such image sections are shown in table 1, and the results on clustering the 25 image sections of the acetowhite region, shown in figure 4(a), using these three features is shown in figure 4(b). Table 2 shows the average area, green and blue values for the same four image sections as in table 1, and figure 4(c) shows the corresponding sections that were found to have mosaic sections. In both figures 4(b) and 4(c), the region shown in color is the region found to contain mosaic patterns.

Table 1: Average area, Cb and Cr values for four image sections

AVERAGE AREA	AVERAGE Cb	AVERAGE Cr	MOSAIC
54.75	116	144	
62.875	119	146	X
64.0625	120	146	X
69.75	118	148	X

Table 2: Average area, green and blue values for four image sections

AVERAGE AREA	AVERAGE G	AVERAGE B	MOSAIC
54.75	144	132	
62.875	130	122	X
64.0625	129	124	X
69.75	125	116	X

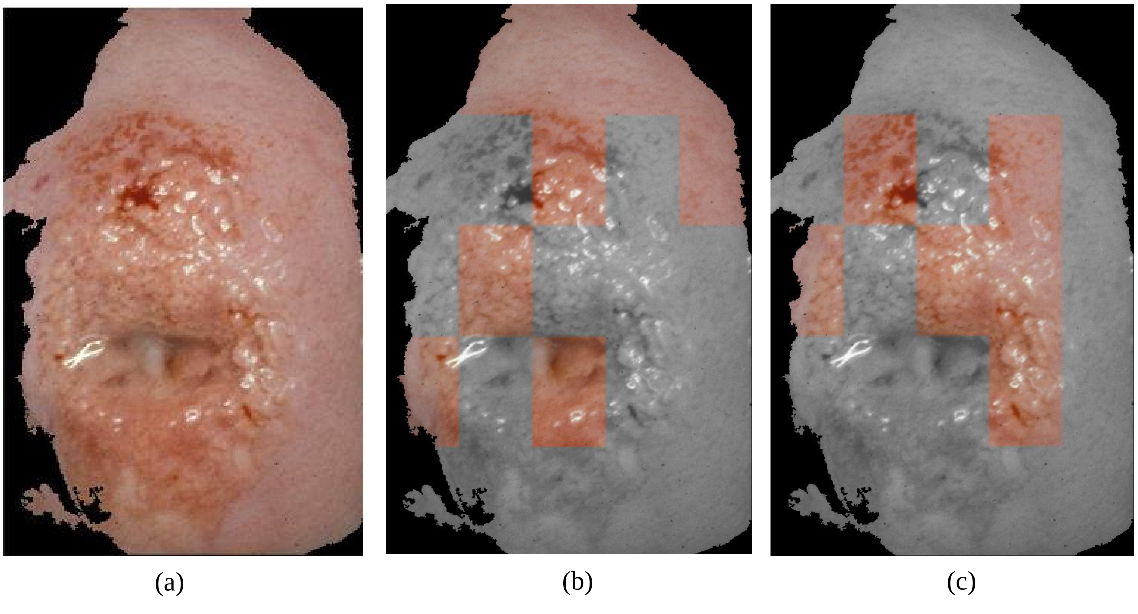
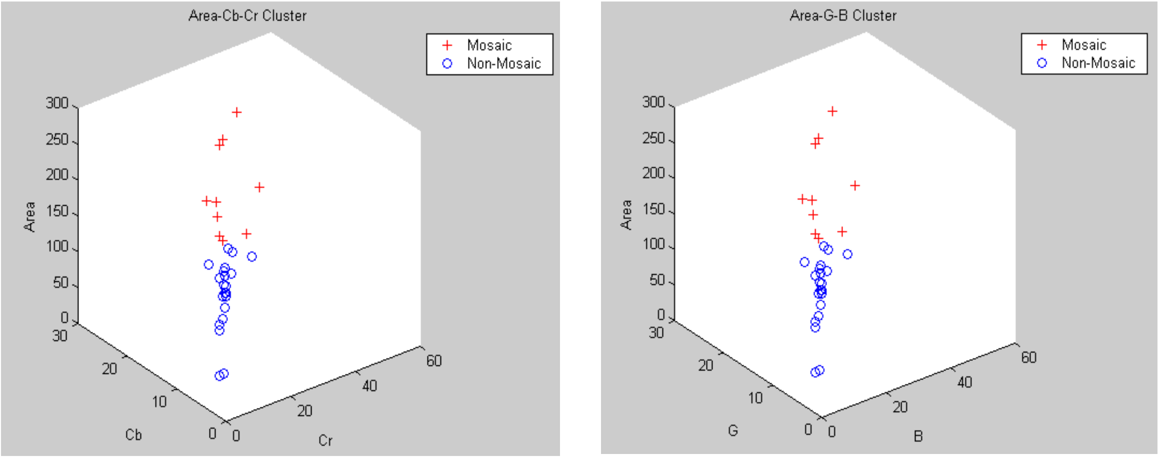


Figure 4: (a) Acetowhite region (b) Regions identified with mosaic patterns using average area, Cb and Cr (c) Regions identified with mosaic patterns using average area, green and blue values



(a) (b)

Figure 5. Average values for mosaic and non-mosaic image sections in 3D space (a) Area-Cb-Cr space (b) Area-green-blue space

14 of the 25 sections of the image were found to have mosaic patterns as indicated in figures 4(b) and (c). Between [area, Cb, Cr] and [area, green, blue] features, all regions containing mosaic vascular patterns are identified. While the current implementation uses image sections of equal size, the size of the sections could be made to change adaptively depending on the position on the image. This would enable using larger image sections in regions where no patterns are expected, like the central part of the image.

Figure 5(a) and (b) show the separation between mosaic and non-mosaic clusters in 3D space. Figure 5(a) plots the average area, Cb and Cr values for image sections from two different images and figure 5(b) shows the average area, green and blue values for image sections from the same images. The separation between the two regions is quite apparent.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a fully automatic method for segmenting the acetowhite section from a cervix image and classifying mosaic and non-mosaic sections in the segmented acetowhite image. The methodologies of our approach were described and results were shown to validate it. While this approach was found to work well under the limited conditions it was tested, this work is by no means complete. There is also a need to extend this study to identify and classify other characteristics of CIN such as punctations and abnormal vasculature, which we hope to address in our future transactions.

We are also investigating on methods to improve the robustness of ROSE and its applicability to both colposcopic and poorer quality cervicographic images. Our current database of images are mostly digitized images from scanned cervigrams. We are collaborating with gynecologic oncologists to obtain the digital images directly from the colposcope so that we can validate our automated segmentation and classification procedures on each type of images and determine if these procedures are equally applicable to poorer quality and less expensive imaging devices such as cervicoscopes for use in resource poor regions.

REFERENCES

1. American Cancer Institute, Cancer Facts and Figures 2005.
2. http://cancer-symptoms.org, Cervical Cancer Information.
3. http://www.nci.nih.gov/, News Article - Treatment Options for Cervical Cancer by Stage.
4. National Cancer Institute, Cancer Facts - The Pap Test: Questions and Answers, http://cis.nci.nih.gov/fact/5_16.htm
5. Women's Health at www.about.com, News article - What is Colposcopy?, http://womenshealth.about.com/cs/cevicalconditions/a/colposcopy.htm.
6. D. G.Ferris. "Cervicography-An adjunct to Papanicolaou screening," *Am Fam Phys.* , Vol. 50, pp. 363-70, 1994.
7. L. Denny, L. Kuhn, A.Pollack, H. Wainwright, T.C. Wright, Jr., " Evaluation of alternative methods of cervical cancer screening for resource –poor settings," *Cancer*, Vol. 89, pp. 826-833, 2000.
8. L.G. Koss, "The Papanicolaou test for cervical cancer detection. A triumph and tragedy," *JAMA*, Vol. 261, pp. 737-774, 1989.
9. Schiffman MH. Arguments against the routine clinical use of currently available HPV screening tests. Contemp OB/GYN. 1990;35:34-46.
10. D. Ferris, "Analysis of digitized cervical images to detect cervical neoplasia", Proc. SPIE Medical Imaging, Vol. 5370, Keynote Speech, San Diego, February 17, 2004.
11. J. Jerenimo, P.E. Castle, R. Herrero, R.D. Burk, M. Schiffman, " HPV testing and Visual Inspection for cervical cancer screening in resource-poor regions" International Journal of Gynecology and Obstretrics , Vol. 83, pp 311-313,2003.
12. Journal of the National Cancer Institute Monographs, Future Directions in Epidemiologic and Preventive Research on Human Papillomaviruses and Cancer, Number 31, 2003.
13. Viara Van Raad, Andrew P. Bradley, "Active Contour Model Based Segmentation of Colposcopy Images of Cervix Uteri Using Gaussian Pyramids," *6th International Symposium on Digital Signal Processing for Communication Systems*, Sydney, Australia, January 2002.
14. Shiri Gordon, Gali Zimmerman, Hayit Greenspan, "Image Segmentation of Uterine Cervix Images for Indexing in PACS," 17th Symposium on Computer-Based Medical Systems, Bethesda, Maryland, June,2004..
15. Shuyu Yang, Jiangling Guo, Philip King, Y. Sriraja, Sunanda Mitra, Brian Nutter, Daron Ferris, Mark Schiffman, Jose Jeronimo, Rodney Long , "A multi-spectral digital cervigram™ analyzer in the wavelet domain for early detection of cervical cancer", SPIE Proc. on Image Processing, **Vol. 5370**, pp 1833-1844, 2004.
16. M. Anderson, J. Jordon, A. Morse, Frank Sharp, A Text and Atlas of Integrated Colposcopy, Mosby Year Book, 1991.
17. B.D. Thackray, A.C. Nelson, "Semi-Automatic Segmentation of Vascular Network Images Using a Rotating Structuring Element (ROSE) with Mathematical Morphology and Dual Feature Thresholding," IEEE Transactions in Medical Imaging, **Vol. 12**, No. 3 September 1993.
18. Qiang Ji, John Engel, Eric Craine, "Texture Analysis for Classification of Cervix Lesions," IEEE Transactions on Medical Imaging, **Vol. 19**, No. 11, November 2000.
19. K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, **Vol. 86**, No. 11, pp. 2210-2239, Nov 1998.
20. N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," IEEE Transaction on Systems, Man and Cybernetics, SMC -9(1), 62-66, 1979.
21. www.mathworks.com, Image Processing Toolbox User's Guide – Converting between Device-Dependant Color Spaces.
22. C. M. Bishop, Neural Networks for Pattern Recognition. Oxford University Press, Oxford, England 1995.