# Technology for Medical Education, Research, and Disease Screening by Exploitation of Biomarkers in a Large Collection of Uterine Cervix Images

L. Rodney Long[a], Sameer Antani[a], Jose Jeronimo[b], Mark Schiffman[b], Mike Bopf[a], Leif Neve[a], Carl Cornwell[a], Scott R. Budihas[a], George R. Thoma[a]
[a]*National Library of Medicine, National Institutes of Health, DHHS,Bethesda, MD 20894*
[b]*National Cancer Institute, National Institutes of Health, DHHS, Bethesda, MD 20894*
*rlong@mail.nih.gov*

## Abstract

*The Communications Engineering Branch of the National Library of Medicine is collaborating with the National Cancer Institute (NCI) in developing applications for medical education, research, and disease screening for precancer detection in the uterine cervix. These applications include (1) expert marking/labeling of tissue regions, (2) Web viewing/ interpretation of histology images, (3) image database/retrieval, and (4) training/testing in clinical image interpretation. Initial NCI studies have been conducted in expert cervicography marking and histology evaluation. We are working toward making cervix images searchable by content-based image retrieval (CBIR). Image pre-processing to remove specular reflection artifacts has achieved 90% success (120 images). Similar results have been obtained for automated location of cervix regions, using Gaussian Mixture Modeling (GMM) with Lab color and one geometric feature. We describe initial classification experiments to discriminate clinically significant tissue, using RGB, HSV, Lab, and YCbCr color models, texture measures, and GMM, Fuzzy C-means, and deterministic annealing algorithms.*

## 1. Introduction

The Communications Engineering Branch (CEB) of the National Library of Medicine (NLM) is collaborating with the National Cancer Institute (NCI) in developing a suite of open source and non-profit applications for the purpose of exploiting extensive longitudinal study data collected on subjects from the United States and in Guanacaste, Costa Rica, a geographic region with relatively elevated rates of cancer of the uterine cervix. Image data collected includes cervicogrpahy (a type of high-definition cervical photograph), Pap test, and histology images. In conventional cervical cancer prevention programs, abnormal cytology (Pap tests) trigger referral to a magnified visual assessment of the cervix following application of vinegar (5% acetic acid), which is called colposcopy. Cervicography is a low-cost alternative to colposcopy that produces similar images. Colposcopists take biopsies based on their assessment of the site of most significant disease. The resultant biopsies are used to guide treatment. While biopsy and cytology slides are saved and can be shared for research and teaching, colposcopy has not lent itself to rigorous research. The use of stored digital images is expected to make an impact on research and education in the use of cervicographic and colposcopic images for the study and prevention of uterine cancer. It has been remarked by one expert in the field of gynecology that colposcopy research has lagged behind other fields that have taken advantage of advances in computerization [1].

> Most colposcopy research has been conducted at single institutions or small consortia, limiting generalizability and scope. The ASCCP [American Society for Colposcopy and Cervical Pathology], NCI, and NLM have developed a panel of colposcopists who will be assessing research questions in colposcopy through the Internet by marking up [thousands of cervicographic] images using NIH-based servers and statistical techniques. This collaboration should allow us to address issues important to all colposcopists, including how closely the range of colposcopic findings correlates with high-grade disease and how interobserver and intraobserver variability impacts the accuracy of colposcopic assessment. [2]

Four related applications are in development: (1) the Boundary Marking Tool, which allows expert region-marking and labeling of significant visual signs or biomarkers of risk in the cervigrams; (2) the Virtual Microscope, which allows viewing and interpretation of histology images; (3) the Multimedia Database Tool, which provides an integrated database framework for querying and viewing all of the collected data and images; and (4) the Teaching Tool, which is expected to serve as a major tool for training and education of cervicography and colposcopic image interpretation. In addition to providing the conventional functions of archiving, retrieval and data display capabilities across the Web, CEB is actively

engaged in research toward the capability to index and retrieve the contents of the cervicography images directly by image content. This paper provides an overview of the applications under development, their interrelationships and intended use, and a description of CBIR work on the images.

(In this paper we use the term *colposcopy image* to mean an image acquired from the endoscopic instrument, the colposcope; and the term *cervicography image* to mean a cervix image acquired by other means, such as the 35 mm still camera protocol used in the Guanacaste Project. We note that "cervicography" is sometimes used inclusively in the medical literature to refer to all types of photographic image acquisition for the cervix.)

## 2. NCI uterine cervix collection

There are two major studies funded by NCI that have included longitudinal cervical image collection. The first is the ASCUS-LSIL Triage Study (ALTS), a 2-year longitudinal study of 5,000 women with minor cervical cytologic abnormalities that yielded 40,000 cervicographic images. We will not discuss ALTS here. Instead, we will describe a very similar, screening project called the Guanacaste Project [3]. The Guanacaste Project is an intensive, population-based cohort study of human papillomavirus (HPV) infection and cervical neoplasia among 10,000 women in Guanacaste, Costa Rica, where the rates of cervical cancer are perennially high. State-of-the-art visual, microscopic, and molecular screening tests are being used to examine the origins of cervical precancer/cancer and to explore viral and host factors that make a geographic region 'high risk'. The Guanacaste study has completed its field phase after seven years of follow-up, and now has spawned a variety of subprojects based on collected specimens, visual images, and outcomes. NCI is examining several potentially important etiologic cofactors, such as chronic inflammation and endogenous hormone levels, which may contribute to cervical cancer risk. Most ambitiously, over 30,000 cervical cell and 30,000 plasma specimens are being tested for HPV DNA and antibodies, respectively, to determine how type-specific HPV DNA types (there are over 40 types of cervical HPV) and antibodies influence outcome. NCI and NLM are collaborating to develop methods to permit exploration of visual aspects of HPV and cervical neoplasia. In etiologic studies NCI will relate the numbers of infecting viral types with numbers and positions of lesions. NCI will be able to follow the topographic progression and regression of lesions. For screening research NCI will be able to use 60,000 digitized uterine cervix images from the Guanacaste Project to optimize and standardize visual screening of the cervix. NLM has the role of developing tools and technologies used in these studies.

## 3. Visual Signs or Biomarkers of Risk in the Cervigrams

Biomarkers of significance in the cervicography images, as used in the field of gynecological oncology for the study and evaluation of precancerous conditions of the uterine cervix are characterized as regions with a complex set of attributes. Color, texture, and relative geometry are predominately useful, while region shape is significantly less so. Regions are frequently amorphous, or, for a few region classes, exhibit a shape which may be only very approximately modeled, and even in these cases, the model may be image dependent. The overall region of interest in the images is the cervix area, the roughly elliptical region that contains all of the relevant anatomy. Within this region, and usually located approximately centrally, is the os, or opening into the uterus. The os has an approximately circular shape in women who have not borne children, and an elongated, more irregular shape in others. The os is typically surrounded by a region of glandular tissue called the *columnar* region, coarsely textured and frequently exhibiting red and/or white coloration, which is in turn surrounded by *squamous* tissue like the inside of the cheek, which tends to be smoothly textured and pinkish. The boundary between these two regions, the *squamocolumnar junction*, is of particular interest to clinicians and researchers, since it is here that most precancer develops. Like colposcopy, the cervigrams are taken directly after the application of 5% acetic acid to the cervix, since it is known that this treatment elicits a visually transient *acetowhite* (AW) appearance in human papillomavirus (HPV) infected or precancerous regions. HPV infection causes virtually all cases of cervical cancer. The acetowhitened regions are high-interest biomarkers in these images. In Figure 1 the three regions shown correspond, in order of region size, respectively, to (1) the overall cervix area, (2) an acetowhite region, and (3) the os. To date, the identification of these biomarkers has been done manually by experts. Our CBIR research, however, is targeted toward automated or computer-assisted biomarker identification.

## 4. The applications

The main characteristics of the four applications are described below. All are Internet-enabled applications, either with client/server architecture, or browser-based, as noted in the individual descriptions.

### 4.1 Boundary Marking Tool

The Boundary Marking Tool (BMT) [4], shown in Figure 1, provides capability to manually draw regions on the cervicography image and to record region labels and expert interpretative information. Regions that may be marked which correspond to tissue types or anatomical features are *acetowhite lesions, invasive cancer,* squamous metaplasia, *Nabothian cysts*, *cervical borders*, *os*, and *polyps*. In addition, the squamocolumnar boundary may be marked, as well as two frequently-obscuring features: blood and mucus. Detailed labeling may be recorded for some of these features. For example, for the acetowhite lesions, the expert may classify the lesion boundary shape characteristics using a standard Reid scale, may classify the color of the lesion, and may record whether certain detail features (punctation, mosaicism, vasculature) are present. The BMT is a mature tool, primarily for data collection, that has already supported studies for NCI researchers, including one published result [5]. A recent NCI data collection with the BMT used cervicography from 939 women and 20 expert colposcopist evaluators at geographically-distributed sites. Each evaluator marked cervix boundaries and acetowhite lesions on the images and provided a clinical diagnosis, ranging from *normal*, through *low-* and *high-grade lesion*, to *invasive cancer*. Multiple studies are under way on this data, including assessment of reproducibility of colposcopic diagnosis, accuracy of the visual evaluation, and visual patterns of cervices of HPV-infected and non-HPV-infected women. Future BMT studies will investigate patterns of appearance and disappearance of precancerous lesions, and inter-observer agreement on biopsy placement. The BMT is designed as a Java client application which interfaces to a server MySQL database; tunneling software (JDBTunnel) is currently used to allow the client to communicate to the database by using only HTTP messages to the Web server. This avoids communications problems frequently encountered when users deploy the BMT client behind firewalls at their local sites. In further development of this tool, we plan to replace the tunneling software with a servlet architecture to avoid the firewall problem without this commercial software dependency.

### 4.2 Virtual Microscope

The Virtual Microscope (VM) provides capability to view histology images and to record expert interpretations. The current, operational prototype is shown in Figure 2. Common practice in current histology studies by multiple experts is to use physical microscopes and glass slides, with the slides being sequentially shipped from one expert to another for interpretation. The VM will allow simultaneous viewing and interpretation of histology by multiple experts at geographically-distributed sites. With the VM a study administrator may create a set of research questions and identify associated regions on histology images; the VM then presents these questions, with a display of the associated regions, to experts who are participating as study observers, and records their answers in a server database. The very large size of histology images, ranging into the tens of gigapixels, requires the VM to adopt the method of viewing the images as tiles which are dynamically assembled into the current view panned by the user. Prototypes of the VM have been developed as a browser-based application based on Zoomify [6] tiling technology incorporated in a Web server system with PHP and Java servlets. Figure 2 shows the screen from one of these prototypes. A key NCI study underway is evaluating the reliability of virtual microscope technology in diagnosis of pre-malignant uterine cervix disease. In this study digitized slides from 600 patients are viewed by five expert observers, who record a diagnosis ranging from *normal*, through *CIN 1/2/3*, to *invasive cancer*. The study results will be compared to a study previously carried out with the same data, using conventional microscope/glass slide protocol. The final VM system is being designed with Java support for image tiling and handling. The VM is primarily a data collection tool.

### 4.3 Multimedia Database Tool

The Multimedia Database Tool (MDT), shown in Figure 3, is the central database tool for accessing the Guanacaste Project data. The MDT is a follow-on program to the Web-based Medical Information Retrieval System [7], with which NLM has been distributing spine x-ray images and health survey data for several years, and provides the capability to query on any of the text data in the MDT Guanacaste Project database, and retrieve not only text, but associated images. The MDT will allow query of the central repository of all of the cervicography images (including those marked by the BMT), histology images, Pap test images, and other images associated with the Guanacaste Project (or ALTS or other similar projects). One area of design emphasis has been on supporting a patient-centric view, as requested by NCI medical collaborators, that will enable all data related to a particular patient, including both text and images, to be navigated in a streamlined manner: for a particular patient, users will be able to view and move among different image types (cervicography, histology, Pap test) on a multi-view display, as well as to dwell on a particular image type and drill through a stack of images for that image type.

The MDT is at an intermediate level of development and is frequently used for internal and conference demonstrations by NLM and NCI. The MDT is a three-tier application consisting of a Java desktop client and servlet, and a MySQL database server. It is used primarily for data dissemination.

### 4.4 Teaching Tool

The Teaching Tool (TT), shown in Figure 4, provides training and teaching in the interpretation of cervicography images for development of precancer. It is intended for use in training experts in the use of cervicography and colposcopy images for screening patients for pre-cancerous conditions. The TT allows a study administrator to create training materials, in the form of images and related questions, for which immediate feedback may be provided; or certification examinations, which also present images and related questions, but with student responses being collected, scored, and sent to the study administrator, with statistical summaries of results provided by the tool. An example question from a certification exam might present images of both a Pap test and cervicography and ask, "Based on your overall impression, what is the worst diagnosis?", where multiple choice answers ranging from *HPV infection* to *carcinoma* are provided. Other questions may present histology images showing two biopsy results and similarly ask for a multiple-choice diagnosis, then ask in a follow-up question for clinical management options. The TT has completed the first phase of development as a PHP-driven Web browser application. This version was reviewed by NCI and ASCCP experts who produced requirements for the second phase, which is expected to be deployed for actual training and testing. The TT serves both data dissemination (by training) and data collection (by testing) functions.

### 4.5 How the applications relate

The four applications are coupled through the image data. Cervicography images marked and interpreted with the BMT are stored in the MDT database and used in the Teaching Tool. Histology images interpreted to consensus in the VM will be available, along with the interpretations, in the MDT.

## 5. Indexing for CBIR

NLM is working with collaborating research groups toward segmenting and labeling the important tissue or anatomical regions in the cervicography images with automatic or computer-assisted methods. This work includes algorithms to compensate for irregular illumination effects, or confounding factors such as intense illumination from camera flash. NCI experts have provided truth sets of labeled tissue regions for evaluating segmentation performance.

One of these efforts, led by Greenspan [7], has concentrated on segmentation of the images into three regions, corresponding to columnar epithelium, acetowhite lesions, and squamous epithelium. The method is a 5-step process that (1) detects the overall cervix region by modeling the image as a Gaussian Mixture Model (GMM) with two densities—one modeling the cervix region, the other non-cervix region; features used in the model are the *a* color channel in Lab color space and the distance d from the center of the image; (2) detects bright camera flash areas—specular reflections—by a multistep process that includes thresholding for areas of high brightness and low saturation, filtering the pixels in these areas to select only those that are in the vicinity of high gradients, and applying a 2-density GMM using HSV saturation and value as features to distinguish specular reflection (high value, low saturation) from non-specular reflection; (3) fills in the areas that were identified as specular reflections and removed in the previous step; the image color in the vicinity of the specular reflection areas is propagated into the regions removed as specular reflections; (4) segments the pre-processed image into columnar epithelium (CE) and non-columnar epithelium by applying a 2-density GMM where the model features are both texture-based (specifically, polarity and "texture-contrast" are used; see [8] for specific formulation); (5) finally, segments acetowhite lesions and squamous epithelium (SE), within the non-columnar cervix area, applying a 4-density GMM model where Lab color is used as the model feature. For a 120-image truth set manually created by one NCI expert, the detection of the overall cervix region has been found to be successful in over 90% of the trials (see Figure 5 for an example detection), and similar results have been achieved for detection and removal of specular reflections. The performance on segmentation of the columnar, squamous, and acetowhite tissues have not been similarly quantified at this time, although the most successful classification has been achieved for the columnar epithelium. The segmentation of acetowhite lesions includes the truth regions in most cases, but there are some known cases of over-segmentation.

Another segmentation research effort, led by Mitra [9], has concentrated on the AW region and its subregion characteristics. In this work, the cervix region is manually cropped from the image, rather than automatically detected. Then specular reflections are removed in the images by a thresholding/averaging technique: pixels with R, G, and B values above a threshold are set to zero; then these 0-pixels are replaced with the average pixel value in an NxN neighborhood; the image is then converted to

grayscale, and a binarizing threshold is computed by the Otsu method [10], which maximizes the separation between interclass and intraclass grayscale variance; in a small number of test cases, this threshold has been found to separate the image into two regions, one of which contains the AW tissue; application of mathematical erosion operations to this region results in a clean mask that is used to identify a central region within the image that contains the AW tissue (and other tissue as well). In addition, Mitra has applied Fuzzy C-means and deterministic annealing methods to segment the AW regions. This research has included methods for detecting and classifying the "tiling texture pattern" called *mosaicism* and the dot-like pattern called *punctation*. These patterns occur within the AW regions and have diagnostic significance. One approach incorporates morphological operations, thresholding, and skeletonizing on AW regions to yield representations of vascular structure in regions both with and without mosaicism. In tests on 25 mosaic regions and 23 non-mosaic regions classified by a medical expert in 11 images, mosaic-bearing regions were automatically distinguished from non-mosaic regions by density of edge pixels in skeletonized images. This edge pixel density corresponds to inter-capillary distance in the vasculature: in mosaic regions, the inter-capillary distance is greater, hence the density of edge pixels is less. See Figure 6. Y color channel in YCbCr color space is being used to further enhance the classification of regions into mosaic/non-mosaic. A second approach is to apply texton [11] filtering to model texture in the cervicography. From ten selected images, 100x100 pixel samples of five tissue types (SE, CE, AW, AW/mosaic, AW/punctated) were taken, and each sample was converted to grayscale and processed by a 48-filter bank including filters sensitive to scale, orientation, and center surround. For each of the five textures, the filter response vectors were clustered using the K-means (K=100) algorithm yielding, for each tissue type, a texture model consisting of a 100-bin histogram. Initial tests have shown distinctive texton histograms for AW/mosaic versus AW/punctuated tissue, even when the original graylevel histograms for these images are very similar [12], as illustrated in Figure 7.

## 6. Summary

The National Library of Medicine and the National Cancer Institute are engaged in research and development to make publicly available clinical and image data related pre-cancer in the uterine cervix for medical education, biomedical research, and screening/prevention programs. This effort involves both (1) system-building of applications to annotate, archive, display, and disseminate research and training data, and to collect certification examination results (see Table 1); and (2) research into content-based image retrieval methods for the automatic or computer-assisted indexing of visual signs or biomarkers of risk in cervicographic images.

**Acknowledgements**

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

**References**

1. Jeronimo J, Schiffman M. Colposcopy at a crossroads. *Am J Obstet Gynecol*. (In press).
2. Massad LS. American Society for Colposcopy and Cervical Pathology and the National Institutes of Health Explore Research Collaboration, *Journal of Lower Genital Tract Disease*, 10(1), Jan 2006, 1-2.
3. Herrero R, Schiffman MH, et al. Design and methods of a population-based natural history study of cervical neoplasia in a rural province of Costa Rica: the Guanacaste Project. *Pan American Journal of Public Health*, 1(15), 1997, 362-375.
4. Jeronimo J et al. Digital Tools for Collecting Data from Cervigrams for Research and Training in Colposcopy, *J Low Gen Tract Disease*, 10(1), Jan 2006, 16-25.
5. Castle PE, Jeronimo J, Schiffman M, et al. Age-related changes of the cervix influence human papillomavirus type distribution, *Cancer Research 2006*, 66(2), Jan. 15, 2006, 1218-1224.
6. http://www.zoomify.com
7. Long LR, Pillemer SR, et al. WebMIRS: Web-based Medical Information Retrieval System. *Proceedings of SPIE Storage and Retrieval for Image and Video Databases VI*, **3312**, San Jose, CA, January 24-30, 1998, 392-403.
8. Gordon S, Zimmerman G, et al. Content Analysis of Uterine Cervix Images: Initial Steps Towards Content Based Indexing and Retrieval of Cervigrams. *Proceedings of SPIE Medical Imaging*, **6144**, San Diego, CA, February 11-16, 2006.
9. Mitra S, Nutter B, Yang S, Karp T. Hybrid vector scalar (HVSQ) compression for multiple image classes. Tech rep, NLM Comm Engineering Branch, Jan 24, 2006.
10. Otsu N. A threshold selection method from gray-level histograms, *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-8, no. 1, 1979, 62-66.
11. Leung T and Jitendra M. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1), 2001, 29-44.
12. Mitra S et al.. JPEG2000 capabilities. Tech rep, NLM Comm Engineering Branch, March 10, 2006.

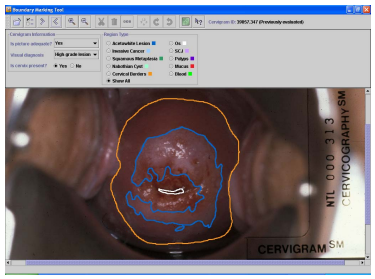


**Figure 1. Boundary Marking Tool, showing three marked regions, in order of decreasing size: (1) overall cervix area, (2) acetowhitened region, (3) os**

**Figure 2. Virtual Microscope (prototype)**

**Figure 3. Multimedia Database Tool**

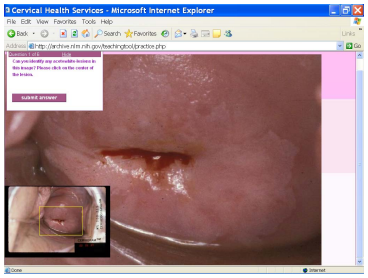


**Figure 4. Teaching Tool**

**Figure 5. Cervix region marked by expert (smaller boundary) and by algorithm**

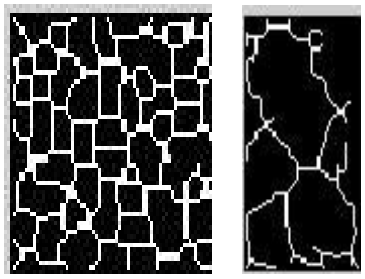

**Figure 6. Left: vascular pattern in non-mosaic region; right: vascular pattern in mosaic region**

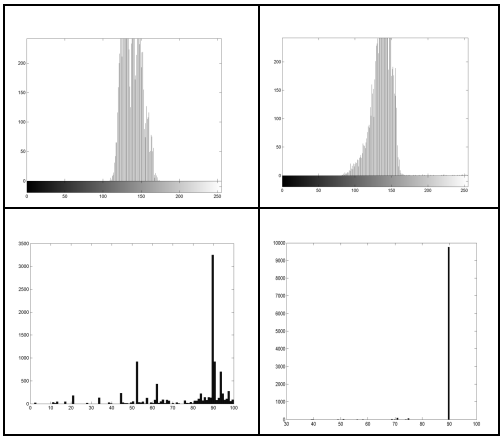

**Figure 7. Texton processing result. Top: grayscale histogram for unprocessed mosaic (left) and punctuated regions. Bottom: 100-bin texton histograms for same regions.**

| Application | Function | Image type |
|---|---|---|
| Boundary Marking Tool | Region labeling by expert | Cervicography, colposcopy |
| Virtual Microscope | Histology image studies | General histology |
| Multimedia DB Tool | Text/image DB access | Cerv., colpo., hist., PAP |
| Teaching Tool | Training/Certification exams | Cervicography, colposcopy |

**Table 1. The four NLM/NCI applications.**