

Content Analysis of Uterine Cervix Images: Initial Steps Towards Content Based Indexing and Retrieval of Cervigrams

Shiri Gordon ^a, Gali Zimmerman ^a, Rodney Long ^b, Sameer Antani ^b, Jose Jeronimo ^c and Hayit Greenspan ^a

^a Tel Aviv University, Tel-Aviv 69978, Israel

^b National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

^c National Cancer Institute, National Institutes of Health, Bethesda, MD 20852, USA

ABSTRACT

This work is motivated by the need for visual information extraction and management in the growing field of medical image archives. In particular the work focuses on a unique medical repository of digital cervicographic images (“Cervigrams”) collected by the National Cancer Institute (NCI) in a longitudinal multi-year study carried out in Guanacaste, Costa Rica. NCI together with the National Library of Medicine (NLM) is developing a unique Web-based database of the digitized cervix images to study the evolution of lesions related to cervical cancer. Such a database requires specific tools that can analyze the cervigram content and represent it in a way that can be efficiently searched and compared. We present a multi-step scheme for segmenting and labeling regions of medical and anatomical interest within the cervigram, utilizing statistical tools and adequate features. The multi-step structure is motivated by the large diversity of the images within the database. The algorithm identifies the cervix region within the image. It then separates the cervix region into three main tissue types: the columnar epithelium (CE), the squamous epithelium (SE), and the acetowhite (AW), which is visible for a short time following the application of acetic acid. The algorithm is developed and tested on a subset of 120 cervigrams that were manually labeled by NCI experts. Initial segmentation results are presented and evaluated.

Keywords: medical image analysis; cervical cancer; cervicography images; colposcopic image; content-based indexing; image segmentation

1. INTRODUCTION

Cervical cancer, the second most common cancer affecting women worldwide and the most common in developing countries, can be cured in almost all patients, if detected by high quality repeated Pap screening, and treated. However, cervical cancer incidence and mortality remain high in resource-poor regions, where high-quality Pap screening programs often cannot be maintained because of inherent complexity and cost. An alternative method of cervical cancer screening, termed cervicography, uses visual testing based on color change of cervix tissues when exposed to acetic acid. This inexpensive method helps to detect abnormal cells that turn white (“acetowhite”) following the application of acetic acid.¹ When additional screening techniques are available, cervicography may be used at the initial examination level, and patients with cervicographic indicators of concern are then referred to colposcopic and/or Pap smear screening.

The National Cancer Institute has collected an extensive set of biomedical information related to the occurrence and evolution of uterine cervical cancer in a longitudinal multi-year study carried out in Guanacaste, Costa Rica. The Guanacaste Project is an intensive, population-based cohort study of human papillomavirus (HPV) infection and cervical neoplasia among 10,000 women in Guanacaste, where the rates of cervical cancer are perennially high. State-of-the-art visual, microscopic, and molecular screening tests are being used to examine

Address all correspondence to H. Greenspan, E-mail: hayit@eng.tau.ac.il

Copyright 2006 Society of Photo-Optical Instrumentation Engineers. This paper will be published in SPIE Medical Imaging Symposium and is made available as an electronic reprint with permission of SPIE. One print or electronic copy may be made for personal use only. Systematic or multiple production, distribution to multiple locations via electronic or other means, duplications of any material in this paper for fee or commercial purpose, or modification of the content of the paper are prohibited

the origins of cervical precancer/cancer and to explore which factors make a geographic region “high risk”. The Guanacaste study has completed its field phase after 7 years of follow-up, and now has changed into a variety of subprojects based on collected specimens, visual images, and outcomes. Data collected included patient age, sexual/reproductive history, laboratory test results, including Pap smear and cytology, and 60,000 cervicographic images in the form of 35 mm color slides, as well as medical classifications for the cervigrams into diagnostic categories.²⁻⁴ Because of the similarity of cervicographic images to colposcopy, NCI and the American Society for Colposcopy and Cervical Pathology (ASCCP) plan to use these images for the training and education of colposcopic practitioners. A major long-term objective is to develop a unique Web-based database of digitized cervix images for investigating the role of HPV in the evolution of cervical cancer and its intraepithelial precursor lesions in women. The NCI/NLM cervigram database contains visual as well as textual data. Unless manually labelled and annotated by a medical expert, the visual data is inaccessible using simple text-based searches. Therefore, the capability to automatically extract this data is highly desirable.

This work is part of an ongoing effort towards the creation of a content-based image retrieval (CBIR) system for the cervicographic images. Content-based image retrieval seeks to extend text-based indexing in a powerful way: to index the images with descriptors derived from the image data itself, in as automated fashion as possible. Effective visual information management, including image coding for efficient communication or storage, or image understanding for image database queries and retrieval, is a topic of great value and research interest.⁵⁻⁷ Content-based indexing and retrieval based on the information contained in medical images is expected to have a great impact on medical image databases.⁸

A few systems have been recently introduced for medical image retrieval. These include text-based systems⁹ as well as the following: ImageMatch of MD Online Inc., the National Medical Practice Knowledge Bank¹⁰ for neuroradiology in computed tomography (CT), the Automatic Search and Selection Engine with Retrieval Tools (ASSERT), the Image Management Environment (IME) for high resolution CTs of the lung,^{11,12} and the Generic Multimedia Indexing (GEMINI) for mammography.¹³ More systems are presented in a recently published review on medical CBIR systems.¹⁴ The current research is the first attempt to create such a system for cervicographic images.

Initial studies can be found on the analysis of individual cervigram images, or the higher-resolution colposcopic images. Most of these studies require the user to mark regions of interest on various cervix tissues.¹⁵⁻¹⁷ Features such as color,¹⁶ texture,¹⁷ and shape¹⁵ are then extracted automatically. Based on these features, the different regions are classified into different cervix tissues using various classifiers, such as neural networks¹⁵ or the minimum distance classifier.¹⁷ Several more recent works have started to address the task of automatic segmentation of specific regions within the cervix. Van Raad¹⁸ performs segmentation of the Transformation Zone (TZ), based on its local frequency content. King et al.¹⁹ perform segmentation of the cervix area within a cervigram based on its color properties. The cervix boundary is then used for cervix registration. Gordon et al.²⁰ introduce initial investigation of adequate feature-spaces for automatic cervigram segmentation. With respect to a CBIR system: preliminary segmentation efforts for the cervigrams within the NCI/NLM database were recently introduced by Zimmerman et al.²¹ and by Srinivasan et al.²² Currently, no study exists that provides a complete analysis of the cervigram images for the content-retrieval task.

Cervigrams contain complex and confusing lesion patterns, and their automatic analysis is a challenging task (for an example cervigram see Figure 1). First, there is a need to cope with artifacts that are generated during the cervigram acquisition process (caused by the strong flash that is used to achieve good illumination, and the convex shape of the cervix). These artifacts include strong shading that cause an inhomogeneous appearance within the tissues and specular reflections artifacts that interfere with the tissues segmentation. Second, a large and complex variability is present within the NCI/NLM archive, and the system should be able to cope with it: The image acquisition setup is not constant. The viewing angle varies strongly between the images; thus, the cervix region differs in intensity and shape across the images. In addition, the physical scene that is imaged has intrinsic variability. For example, in different patients the cervix is not the same size, and additional non-cervix tissues or medical instruments may appear in the image. A third significant difficulty is the variability of content within the images. Not all cervix-tissue types are always present, and there is no prior knowledge regarding the presence of acetowhite lesions. Finally, the narrow dynamic range of the colors, the lack of clear boundaries

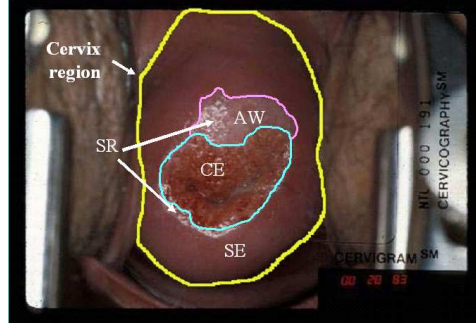


Figure 1. An example cervigram; marked are the cervix region, the different tissues of interest: the columnar epithelium (CE), the squamous epithelium (SE), the acetowhite (AW) and the specular reflection artifacts (SR).

between tissue regions, and the lack of other consistently recognizable landmarks within the images, generate image-understanding challenges that require adequate computational analysis tools.

2. CERVIGRAM SEGMENTATION

The proposed segmentation scheme is an automatic multi-step process, where each step targets a specific region within the cervigram (Figure 2). The algorithm uses unsupervised clustering via Gaussian Mixture Modeling (GMM) in conjunction with feature sets that are most suitable for each step. A brief description of the various steps is given in Sections 2.1-2.4.

The distribution of a d -dimensional random variable is a mixture of k Gaussians if its density function is :

$$f(y) = \sum_{j=1}^k \alpha_j \frac{1}{\sqrt{(2\pi)^d |\Sigma_j|}} \exp\left\{-\frac{1}{2}(y - \mu_j)^T \Sigma_j^{-1} (y - \mu_j)\right\}, \quad (1)$$

where d is the feature space dimension, α_j are the probabilities of the occurrence of each Gaussian, and μ_j, Σ_j are the mean and the covariance matrix of each Gaussian cluster respectively. The Expectation-Maximization (EM) algorithm²³ is used to determine the maximum likelihood parameters of a mixture of k Gaussians in the selected feature space. An immediate transition is possible between GMM and probabilistic labelling of a feature vector y , where each feature vector is assigned to the most probable Gaussian cluster in the learned GMM:

$$label(y) = \arg \max_j \alpha_j f(y | \mu_j, \Sigma_j) \quad j = 1 \dots k. \quad (2)$$

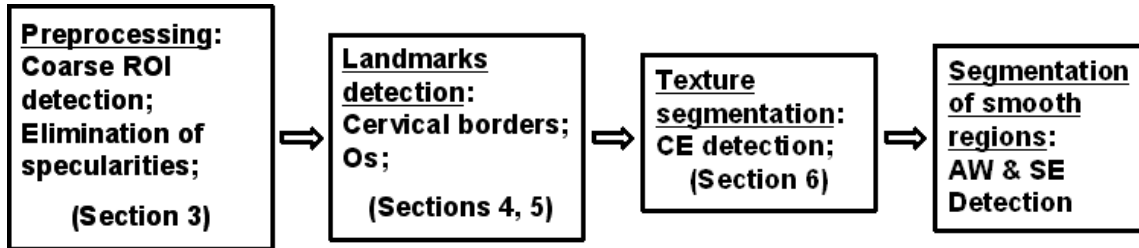


Figure 2. Block diagram of the multi-step segmentation scheme.

2.1. AUTOMATIC EXTRACTION OF THE CERVIX REGION

The cervix region occupies about half of the cervigram image. Other parts of the image contain irrelevant information, such as equipment, frames, text, and non-cervix tissues (Figure 1). This irrelevant information can confuse the automatic identification of the tissues within the cervix. The first step is therefore focusing on the cervical borders, so that we have a geometric bound on the relevant image area. The cervix region is a relatively pink region located around the image center. Two features are utilized for its automatic extraction: the a color channel of the Lab color space (the higher the value of a , the “redder” is the pixel’s color) and d - the distance of a pixel from the image center. The a color channel is initially smoothed in order to eliminate small details. A coarse Region-of-Interest (ROI) is automatically extracted based on the $a - d$ feature space and Gaussian mixture modeling. The cluster that has the lowest $mean(d)$ and the highest $mean(a)$ is identified as the ROI. When the resulting ROI consists of several disjoint areas in the image, the largest one is chosen, and the others are ignored. The image is cropped to include the ROI region, and subsequent steps of the process are performed within it, thus avoiding the confusing patterns and colors that occupy the rest of the image. An example for this process is presented in Figure 3.

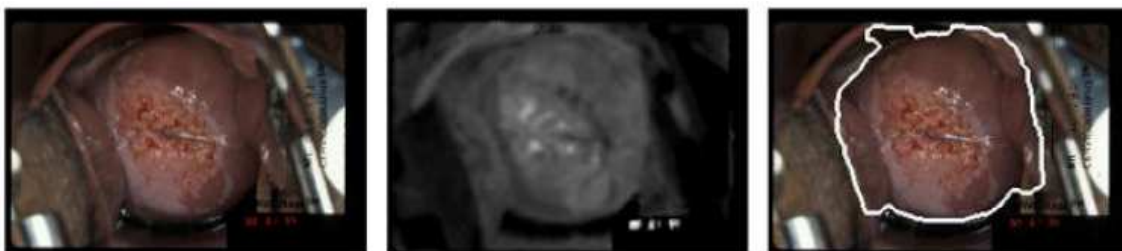


Figure 3. ROI identification: Original cervigram (left); Smoothed a feature (center); ROI detected by the clustering algorithm - marked by a white line (right).

2.2. IDENTIFICATION OF SPECULAR REFLECTIONS

Processing of many of the cervigrams is complicated by the presence of Specular Reflection (SR) artifacts. These artifacts are small and bright regions on the cervix surface, which are generated during the image acquisition process due to the presence of fluids (Figure 1). The SR artifacts provide misleading “tissue” information and interfere with the content analysis of the regions surrounding them. Their modeling and segmentation is difficult, since the number of pixels belonging to SR regions is small relative to the total number of pixels in the image. In order to overcome these difficulties a two-stage process has been recently developed.²⁴ In the first step, candidate SR regions are detected by imposing thresholds of high brightness, I , and low color saturation, S , on the image.²⁵ Following the thresholding process, the candidate SR mask is refined by selecting only the pixels in the vicinity of high gradients. Pixels within the candidate regions are next mapped into a two dimensional $S-V$ feature space, where S and V are Saturation and Value of the standard HSV color space. The extracted 2D feature space is modeled as a mixture of two Gaussians. One represents the brightest, almost white pixels that have lost most of their chromatic information. The other Gaussian represents the less bright pixels that retain some of their original color. Pixels corresponding to the Gaussian with the highest mean intensity are selected as the SR pixels and are removed from the image. Following the SR extraction phase, the generated “holes” are filled to enable further analysis of the image content. A simple filling scheme that eliminates the strong gradients associated with the SR, while preserving the original texture, is currently used. It is based on the observation that the highlights formed on the moist surface of the cervix are very small. The color underneath the highlights is assumed to be nearly constant and similar to the color of the pixels in the immediate surroundings. The color of the surrounding pixels is thus propagated into the now-empty specular regions accordingly.

2.3. TEXTURE SEGMENTATION

The columnar epithelium (CE) appears as a coarse and textured region, with no specific texture pattern. The rest of the image is relatively smooth and non-textured. A multi-scale *texture-contrast* feature is extracted for

its detection. This texture feature describes both the underlying texture parameters and the adequate texture scale.²⁶ The scale is defined as the width of the Gaussian window within which gradient vectors of the image are pooled. The second moment matrix for the vectors within this window, computed about each pixel in the image, can be approximated using:

$$M_\sigma(x, y) = G_\sigma(x, y) * (\nabla I)(\nabla I)^T, \quad (3)$$

where G_σ is a separable binomial approximation to a Gaussian smoothing kernel with variance σ^2 , and (∇I) is the gradient of the image intensity. Two texture descriptors are extracted for each pixel: polarity and texture-contrast. Polarity is a measure of the extent to which the gradient vectors in a certain neighborhood all point in the same direction, defined as:

$$p_\sigma = \frac{|E_+ - E_-|}{E_+ + E_-}, \quad (4)$$

where σ is the scale. The definitions of E_+ and E_- are:

$$\begin{aligned} E_+ &= \sum_{x,y} G_\sigma(x, y) [\nabla I \cdot \hat{n}]_+, \\ E_- &= \sum_{x,y} G_\sigma(x, y) [\nabla I \cdot \hat{n}]_-, \end{aligned} \quad (5)$$

where $[\cdot]_+$ and $[\cdot]_-$ are the rectified positive and negative parts of their argument and \hat{n} is a unit vector perpendicular to ϕ (the direction of principal eigenvector of the second moment matrix, as defined in Equation 3). This feature is used later for the selection of an appropriate texture scale for each pixel in the image. The contrast relates to the energy of the gradients in the vicinity of each pixel as given by Equation 6, where λ_1 and λ_2 are eigenvalues of M_σ ($\lambda_1 \geq \lambda_2$)

$$contrast = 2\sqrt{\lambda_1 + \lambda_2} \quad (6)$$

The process of selecting an appropriate scale is based on the derivative of the polarity with respect to the scale. For each pixel (x, y) the scale is selected as the first value for which the difference between values of polarity at successive scales is less than 2%.

At the end of the texture features extraction phase, each pixel is associated with a contrast feature of the appropriate scale. GMM clustering is performed using this feature to obtain two clusters, thus separating between the smooth and the textured regions within the image. This step is performed on the preprocessed cervix area with filled-in SR regions.

2.4. SEGMENTATION OF SMOOTH REGIONS

Following the detection and extraction of the textured regions, only the smooth regions within the cervix remain. These regions are modeled as a mixture of four Gaussians in the *CIE-Lab* color feature space. The cluster with the highest mean intensity is identified as a candidate cluster to include the AW lesions. Initial results have been recently published by Zimmerman et al.²¹ Additional examples are presented in Section 3. The current process of AW detection is still at its very preliminary stages. It should be improved to support the large variability of AW lesions within the database. Due to illumination effects the AW and the SE tissues often possess very similar colors, and AW lesions are wrongly detected. On the other hand AW lesions located in shaded areas of the image are not detected at all.

3. RESULTS

Each step presented in this work is a significant task in itself, and is therefore evaluated separately. Manual segmentations are available for 120 images, and an initial performance evaluation has been carried out. Figures 4-7 show example results of the different steps, along with the manual segmentations. The detection of ROI is successful in 100% of the images, since it always includes the cervix and excludes most of the irrelevant surroundings. The detection of the specularities was found highly satisfactory by medical experts in 90% of the images.²⁴ SR are effectively removed, thus enabling an improved performance in subsequent stages. The texture detection step effectively separates the columnar epithelium (along with other textured areas). The AW regions are detected but bright areas outside the cervix border are captured as well. Overall, the automated AW segmentation achieved results that are close to the desired segmentation but generated some false positives.



Figure 4. Examples for automatic detection of the cervix region within the cervigram (white contour)

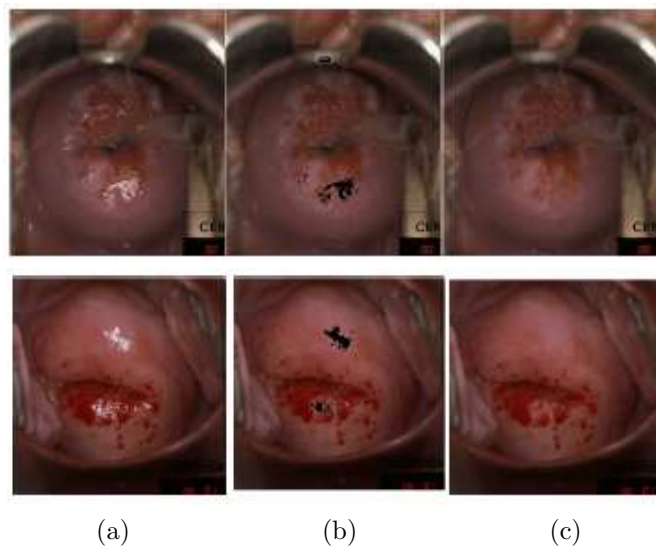


Figure 5. Detection and elimination of specular reflections. (a) Original cervigrams, automatically cropped around the cervix; (b) The detected specularities marked in black; (c) Filled-in images specularities eliminated.

4. DISCUSSION

A limited number of studies, all in very preliminary stages, address the task of automated cervical image analysis. The current work presents an automatic segmentation algorithm for cervigrams with a focus on three tissue types. The squamous epithelium, the columnar epithelium, and the acetowhite lesions. The AW regions are of particular interest since the detection and estimates of their size are of clinical significance. The tasks of detection of the cervix area, the columnar epithelium and the acetowhite regions, are very challenging due to the large diversity

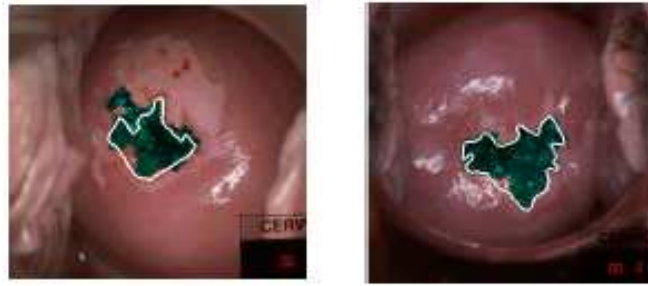


Figure 6. Detection of CE region. Automatically detected regions - green. Manual detection - white.

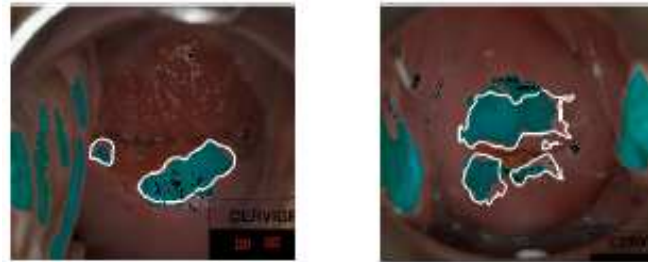


Figure 7. Detection of AW. Automatically detected regions - green. Manual detection contours - white. SR pixels black.

of the cervigram images within the database, and the different artifacts present in the cervigrams. Various problems were addressed within the scope of this work that arise from the real world clinical environment, the most important being the issue of specular reflections. This severe artifact was explicitly investigated and successfully handled. The presented method for SR elimination was developed and optimized specifically for cervigram images since the classical methods, that assume relatively smooth dielectric objects, were found inappropriate. The results underwent initial validation by medical experts and were evaluated as satisfactory. In this contribution we show the ability to detect regions of anatomical and medical interest (the cervix area, CE, AW). Initial segmentation results are promising, and methodical validation by medical experts is currently underway. These results provide a proof of concept for future content-based indexing and retrieval of cervigrams in a Web-based data base. Future work includes further refinement of the cervix region detection, shading correction, using additional features for the CE and the AW detection and the incorporation of higher-level knowledge in order to reduce occurrence of false positives.

5. ACKNOWLEDGEMENTS

This work was supported in part by Contract Number 467-MZ-401687 from the National Library of Medicine, National Institutes of Health, Bethesda, MD, USA.

REFERENCES

1. J. Jeronimo, P. E. Castle, R. Herrero, R. D. Burk, and M. Schiffman, "HPV testing and visual inspection for cervical cancer screening in resource-poor regions," *International Journal of Gynecology and Obstetrics* **83**, pp. 311–313, 2003.
2. R. Herrero, M. H. Schiffman, C. Bratti, A. Hildesheim, I. Balmaceda, M. E. Sherman, M. Greenberg, F. Cardenas, V. Gomez, K. Helgesen, J. Morales, M. Hutchinson, L. Mango, M. Alfaro, N. W. Potischman, S. Wacholder, C. Swanson, and L. A. Brinton, "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), pp. 362–375, 1997.

3. R. Herrero, A. Hildesheim, C. Bratti, M. E. Sherman, M. Hutchinson, J. Morales, I. Balmaceda, M. D. Greenberg, M. Alfaro, R. D. Burk, S. Wacholder, M. Plummer, and M. Schiffman, "Population-based study of human papillomavirus infection and cervical neoplasia in rural Costa Rica," *Journal of the National Cancer Institute* **92**(6), pp. 464–474, 2000.
4. M. Schiffman and P. Castle, "Human papillomavirus," *Archives of Pathological Laboratory Medicine* **127**, pp. 930–934, 2003.
5. S. F. Chang, J. R. Smith, and A. Beigi, "Visual information retrieval from large distributed on-line repositories," *Communications of ACM* **40**(12), pp. 63–71, 1997.
6. H. Stone, "Image libraries and the internet," *IEEE communications Magazine* **37**(1), pp. 99–106, 1999.
7. A. Gupta, S. Santini, and R. Jain, "In search of information in visual media," *Communication of ACM* **40**(12), pp. 35–42, 1997.
8. H. D. Tagare, C. C. Jaffe, and J. Duncan, "Medical image databases - a content-based retrieval approach," *Journal of the American Medical Informatics Association* **4**, pp. 184–198, 1997.
9. L. R. Long, S. R. Pillemer, R. C. Lawrence, G. H. Goh, L. Neve, and G. R. Thoma, "Webmirs- web-based medical information retrieval system," in *Proc. of SPIE*, **3312**, pp. 392–403, 1998.
10. Y. Liu, W. E. Rothfus, and T. Kanade, "Content-based 3D neuroradiologic image retrieval - preliminary results," in *Technical Report CMU-RI-TR-98-04*, (Carnegie Mellon University, Pittsburgh, PA), 1998.
11. C. R. Shyu, C. E. Brodley, A. C. Kak, A. Koska, A. M. Aisen, and L. S. Broderick, "ASSERT - a physician-in-the-loop content-based retrieval system for HRCT image database," *Computer Vision and Image Understanding* **75**, pp. 111–132, 1999.
12. A. F. Abate, M. Nappi, G. Tortora, and M. Tucci, "IME - an image management environment with content-based access," *Image and Vision Computing* **17**, pp. 967–980, 1999.
13. P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas, "Fast and effective retrieval of medical tumor shapes," *IEEE Transactions on Knowledge and Data Engineering* **10**, pp. 889–904, 1998.
14. H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications- clinical benefits and future directions," *International journal of medical informatics* **73**, pp. 1–23, 2004.
15. P. M. Cristoforoni, D. Gerbaldo, A. Perino, R. Piccoli, F. J. Montz, and G. L. Captiano, "Computerized colposcopy: Results of a pilot study and analysis of its clinical relevance," *Obstet. Gynecol* **85**, pp. 1011–1016, 1995.
16. B. W. Pogue, M. A. Mycek, and D. Harper, "Image analysis for discrimination of cervical neoplasia," *Journal of Biomedical Optics* **5**(1), pp. 72–82, 2000.
17. Q. Ji, J. Engel, and E. Craine, "Texture analysis for classification of cervix lesions," *IEEE Transaction on Medical Imaging* **19**(11), pp. 1144–1149, 2000.
18. V. V. Raad, "Frequency space analysis of cervical images using short time Fourier Transform," in *Proc. of the IASTED International Conference of Biomedical Engineering*, **1**, pp. 77–81, (Salzburg, Austria), January 2003.
19. P. King, S. Mitra, and B. Nutter, "An automated, segmentation-based, rigid registration system for cervigram images utilizing simple clustering and active contour techniques," in *Proc. of 17th IEEE Symposium on Computer-Based Medical Systems*, pp. 292–297, 2004.
20. S. Gordon, G. Zimmerman, and H. Greenspan, "Image segmentation of uterine cervix images for indexing in PACS," in *Proc. of 17th IEEE Symposium on Computer-Based Medical Systems*, pp. 298–303, 2004.
21. G. Zimmerman, S. Gordon, and H. Greenspan, "Content-based indexing and retrieval of uterine cervix images," in *Proc. of 23rd IEEE Convention of Electrical and Electronics Engineers in Israel*, pp. 181–185, 2004.
22. Y. Srinivasan, D. Hernes, B. Tulpule, S. Yang, J. Guo, S. Mitra, S. Yagneswaran, B. Nutter, J. Jeronimo, B. Phillips, R. Long, and D. Ferris, "A probabilistic approach to segmentation and classification of neoplasia in uterine cervix images using color and geometric features," in *Proc. of the SPIE Medical Imaging 2005*, **5747**, pp. 995–1003.
23. H. Greenspan, J. Goldberger, and L. Ridel, "A continuous probabilistic framework for image matching," *Journal of Computer Vision and Image Understanding* **84**, pp. 384–406, 2001.

24. G. Zimmerman and H. Greenspan, "Automatic detection of specular reflections in uterine cervix images," in *Proc. of SPIE Medical Imaging*, 2006.
25. T. M. Lehmann and C. Palm, "Color line search for illuminant estimation in real-world scenes," *Optical Society of America* **18**(11), pp. 2679–2691, 2001.
26. C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld:image segmentation using Expectation-Maximization and its application to image querying," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **24**(8), pp. 1026–1038, 2001.