Cervicographic Image Retrieval by Spatial Similarity of Lesions

Zhiyun Xue, L. Rodney Long,
Sameer Antani, George R. Thoma
National Library of Medicine, NIH
{xuez, rlong, santani, gthoma}
@mail.nih.gov

Jose Jeronimo
Reproductive Health, Program for
Appropriate Technology in Healthcare
jjeronimo@path.org

Abstract

The National Library of Medicine has been developing CervigramFinder, a Web-accessible prototype content-based image retrieval (CBIR) system for cervical cancer research, to retrieve cervicographic images from a large collection with respect to visual characteristics of lesion regions. This paper describes current work on retrieving the images based on similarity of spatial location of lesions. The proposed two-level method takes into account the visual characteristics of cervix lesions, as well as spatial information of shape, size, orientation, and distance. The proposed method was evaluated on a data set of 1000 cervicographic images where multiple lesion boundaries as well as associated location information were marked by medical experts. The simplicity and effectiveness of the proposed method was subjectively compared with the Angle Histogram and R-Histogram and was evaluated as better with respect to results ranking.

1. Introduction

Cervical cancer is the second most prevalent malignancy affecting women worldwide, and one of the most common causes of cancer-related mortality in women in developing countries. The vast majority of cervical cancer cases are caused by persistent infections with certain types of the human papillomavirus (HPV) family. The National Cancer Institute (NCI) conducted intensive, population-based, cohort studies in the United States and Costa Rica [1] and collected 100,000 cervicographic images (cervigrams) and some Pap test images and histology slides to investigate the natural history of HPV infection and cervical neoplasia, and compare different cancer screening techniques. A cervigram, shown in Figure 1, is a color picture of the uterine cervix taken with a specially designed optical camera at an approximately one-minute interval after the cervix has been rinsed with a weak solution of acetic acid. This process is called cervicography and is an inexpensive visual screening method similar to colposcopy. The National Library of Medicine (NLM), in collaboration with NCI, has been developing a Web-based system, the Multimedia Database Tool (MDT) [2], to provide a convenient means of data retrieval for the NCI cervigram collection. The MDT has the capability to query the Guanacaste and ALTS text data and to retrieve images as well as text records from the database. Content-based image retrieval (CBIR) techniques are potentially a valuable auxiliary search technique for such databases in the cervical cancer community.

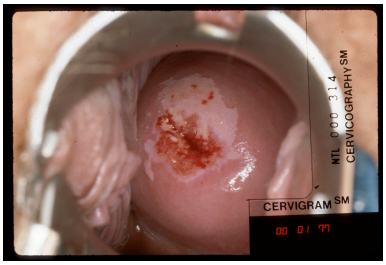

Figure 1. Sample cervigram image

CervigramFinder is a Web-accessible content-based image retrieval (CBIR) prototype system for searching cervigrams by their visual attributes [3]. It operates on an expert-annotated subset of the NCI cervigram data for which the lesion regions-of-interest (ROI) have been manually pre-marked and labeled by NCI medical experts [4]. To operate the system, the user creates a query by marking a region of interest on an image and specifies the types of regions to be searched. The GUI, shown in Figure 2, lets the user

further define the query by specifying the relative importance of each visual feature (color, texture, and size) that is essential in differentiating lesion types and identifying the developing stage of cervical neoplasia. The system then calculates the signature of the query region for the specified features and compares it with the pre-computed signatures in the expert-marked data set. The matching images are returned along with a limited amount of associated text information.

In addition to color, texture, and size attributes, terms such as "at 12 o'clock position" or "in the right upper quadrant" are routinely used to describe the location of a lesion. There is a need to add corresponding functionality into the *CervigramFinder* system. In this paper, we propose a new method, customized to characteristics specific to cervigrams, to index cervix lesions by spatial attributes. The location-based retrieval method operates at two levels: a coarse level and a fine level. At the coarse level, the lesions are coarsely divided into several groups based on their location in the image space. This permits limiting the search space to the vicinity of the query region and makes the search efficient for the next level. At the fine level, detailed features that characterize the relative location of lesions within the query region are used to rank lesions on the target image that belong to the group specified by the location of the lesion in the query.

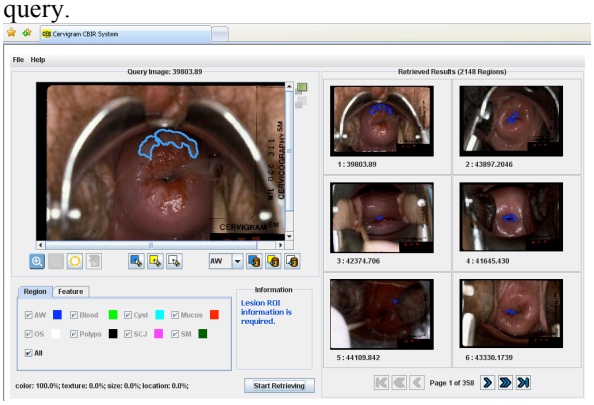

Figure 2. *CervigramFinder* GUI

2. Methods

The location of a region refers to a position or a site occupied by the region with respect to a particular frame of reference. There are three important anatomical features used intuitively by clinicians as references in describing the relative position of cervix lesions: the cervix boundary, the cervix orientation, and the center of the cervical *os* (the opening into the uterus). Conceptually, clinicians project a virtual analog clock face onto the cervix; one reference axis is imagined to lie along the "12 o'clock/6 o'clock" line, and the other reference axis is imagined to be perpendicular to this one. The time markers on this clock face have traditionally served as approximate descriptive locators for items of interest on the cervix. Figure 3 illustrates the position of the *os* and the reference axis which together define a coordinate system and describe the relative position of a lesion on the cervix. We adopt the same terminology as the cervicography medical professionals and refer to this anatomy-fixed axis system simply as "the landmark".

For location indexing, the image space is first divided into four quadrant sectors in the proposed landmark coordinate system. We require a sufficient number of region pixels to lie in a quadrant for it to be considered "occupied". The region is then coarsely identified by these occupied quadrants. The retrieval is then limited to only those regions that occupy the same quadrants as the query region; this constrains the next step, the fine-granularity search, to only these quadrants. This approach is partly inspired by the quadrant description that clinicians employ to coarsely record the lesion location. An additional benefit of this approach is that it does not require the exact cervix boundary information, which is a challenging task. The next step in our approach is to rank the regions by similarity computed with a finer granularity.

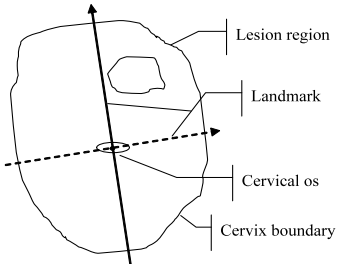

Figure 3. Lesion location characteristics

The location of a region is indirectly related to the shape, orientation, and size of the region. To describe the spatial relations between two objects, some of the earlier approaches in the literature [5] focus on qualitative spatial reasoning and use relational terms such as "left of", "right of", "above", "below", "near", "far", "inside, and "outside". Such approaches are not useful for describing the relative location of regions on the cervix, since all are on the cervix and inside the cervix boundary. Furthermore, such terms, even if used in relation to the *os*, are not readily adaptable for the type of quantitative similarity comparisons and results ranking that we require for CBIR image retrieval. We propose an alternative approach, where the spatial location of a lesion region is characterized by its angle and its distance from the landmark in a polar coordinate system, as shown in Figure 4. This is similar to other methods in the literature, namely:

Angle Histogram [6], *F-Histogram* [7], *R-Histogram* [8] and *R*-Histogram* [9]. Given a region A consisting of N points $(a_i, i = 1,...,N)$ and a reference region B consisting of M points $(b_j, j = 1,...,M)$, the *Angle Histogram* computes the histogram of the angles between each pair of vectors that are determined by the point-pairs (a_i, b_j). The *F-Histogram* generalizes the *Angle Histogram* by handling regions as longitudinal sections instead of points. The *R-Histogram* extends the *Angle Histogram* by incorporating the distances between the points on both boundaries of region A ($a_{c_i}, i = 1,...,K$) and B ($b_{c_j}, j = 1,...L$) and the label indicating whether a_{c_i} (b_{c_j}) is inside B (A). The *R*-Histogram* is an extension to the *R-Histogram*. It generalizes the *R-Histogram* by taking into account all of the pixels in the objects.

Method: The following location features are extracted for the region in the proposed method:

- Normalized angle range of the region

$$\rho_{range} = \frac{\rho_{\max} - \rho_{\min}}{2\pi} \quad (1)$$

where $\rho_{\max}$ and $\rho_{\min}$ are the maximal and minimal angles of the region.

- Normalized radius range of the region

$$r_{range} = \frac{r_{\max} - r_{\min}}{r_{cervix}} \quad (2)$$

where $r_{\max}$ and $r_{\min}$ are the maximal and minimal distances, respectively, from the region to the landmark center; and r_{cervix} is the maximal distance from the cervix boundary to the landmark center.

- Normalized center of the region

$$center = [\rho_c, r_c] = \left[\frac{\sum_i \rho_i}{2\pi N}, \frac{\sum_i r_i}{r_{cervix} N}\right] \quad (3)$$

where, N is the total number of pixels in the region and ρ_i, r_i are the angle and range, respectively, associated with point i.

- Extent of the region

$$e = \frac{N}{(\rho_{\max} - \rho_{\min}) \times (r_{\max} - r_{\min})} \quad (4)$$

The similarity between the locations of two regions is then computed as the Euclidean distance between their feature vectors $f = [\rho_{range}, r_{range}, \rho_c, r_c, e]$.

3. Evaluation and results

The proposed approach was compared with the *Angle Histogram* and *R-Histogram* methods. They were selected based on their general similarity to the proposed method. Since lesions are always inside the cervix boundary, our implementation of the *R-Histogram* method was limited to angle and distance measures, and relative positional information, i.e., the property of points of region A being inside region B was not computed, as was done in the original publication [8]. Further, in the experiments, the *Angle Histogram* was quantized to 36 bins, and the *R-histogram* was quantized to 36 angle bins and 50 distance bins. For both methods, the angles were calculated in the coordinate system defined by the landmark, and histogram intersection was used as the similarity measure.

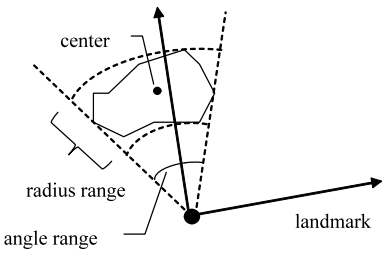

Figure 4. Location features

Data Set: The experiments were performed on a test set of 1000 cervigrams. Each cervigram had been manually marked by one or more experts who drew boundaries for the cervix region and the lesion(s), and positioned landmark axes using the Boundary Marking Tool (BMT) that is an NLM-developed Web-accessible software tool [4]. While the method is, in principle, capable of finding any regions that are situated at a given location, the calculation of similarity ranking for multiple regions, having irregular shapes and sizes, and extending into multiple quadrants, is a challenge that we have yet to address.

Testing: 20 images were arbitrarily selected from the data set for use as experimental queries. The top eight results from each query and for each method were subjectively evaluated by visual inspection by three engineers who have gained a working familiarly with the main visual characteristics of the images through collaboration with NCI medical experts. Evaluation of our results directly by such medical experts is one of the important remaining goals of our project.

Results and Discussion: Figure 5 shows an example of the retrieved results, where, (a) shows the query region, and (b)-(d) shows the top eight returned matching regions resulting from the *Angle Histogram* method, the *R-Histogram* method, and the proposed method, respectively. As expected, the *R-Histogram* and our proposed method appear to outperform the *Angle Histogram* in most cases because the latter does not explicitly consider size and distance information.

The quality of retrieval of our method appears to be better than or comparable to that of the *R-Histogram* method. However, compared to both *R-Histogram* and *Angle Histogram*, our method has two advantages: 1) the computation is faster, which is important for a online real time application (*R-Histogram* is 2D histogram calculating from each pairs of pixels on the region boundaries); 2) strong segmentation of the cervix is not required, which is a big benefit, since it is very hard to automatically segment the exact cervix region due to the high variability in visual appearance of the cervix on the cervigrams. These benefits make our method suitable for use in Web-based CBIR system using color, texture, size, and location characteristics. It can be generalized for spatial region similarity in a polar coordinate space for any application where boundary information is either not available or reliably computable. Results from all 20 queries are available on our Web site[1].

4. Conclusions

Searching an image collection based on spatial properties of regions is an important capability in a medical CBIR system and critical for studying cervicographic images concerned with precancerous lesions. This paper describes a simple but effective 2-tier method for retrieving cervigrams with respect to location and other visual characteristics of cervical lesions. Engineers with expertise in CBIR subjectively evaluated the method on a data subset selected from a multi-national multi-year study tracking HPV types, infections, and their causality on uterine cancer, and compared its performance with two similar histogram-based approaches proposed in the literature for quantitatively representing spatial relations. The proposed method appeared to be effective and efficient. Future work includes medical expert evaluation on a larger data set.

Acknowledgement

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] R. Herrero, M. H. Schiffman, C. Bratti, 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. Rev Panam Salud Publica, No. 1, pp. 362-375, 1997.

[2] L. R. Long, S. Antani, G. R. Thoma. Image informatics at a national research center. Comp Med Imaging and Graphics, 29:171-193, February 2005.

[3] Z. Xue, S. Antani, L. R. Long, et al. A Web-accessible content-based cervicographic image retrieval system. Proc SPIE Medical Imaging, February, 2008.

[4] J. Jeronimo, R. Long, L. Neve, et al. Digital tools for collecting data from cervigrams for research and training in colposcopy. Journal of Lower Genital Tract Disease, 10(1):16-25, January 2006.

[5] A. G. Cohn, S. M. Hazarika. Qualitative spatial representation and reasoning: an overview. Fundamenta Informaticae, 46(1-2):1-29, 2001.

[6] V. N. Gudivada, V.V.Raghavan. Design and evaluation of algorithms for image retrieval by spatial similarity. ACM Trans Information Systems, 13(2):115-144, 1995.

[7] C. Shyu, P. Matsakis. Spatial lesion indexing for medical image databases using force histogram. Proc IEEE Conference on Computer Vision and Pattern Recognition, 2:603-608, 2001.

[8] Y. Wang, F. Makedon. R-histogram: Quantitative representation of spatial relations for similarity-based image retrieval. Proc ACM Multimedia, pp. 323-326, 2003.

[9] Y. Wang, F. Makedon, A. Chakrabarti. R*-histograms: Efficient representation of spatial relations between objects of arbitrary topology. Proc ACM Multimedia, pp. 356-359, 2004.

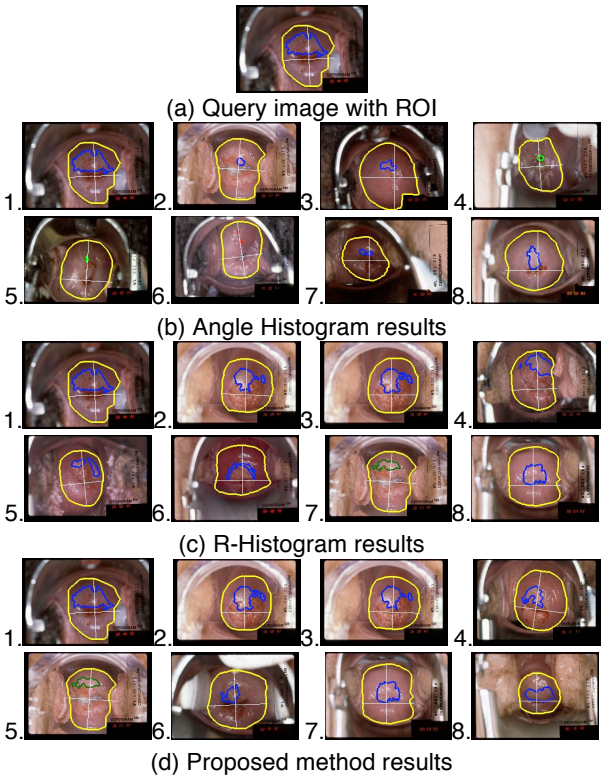

Figure 5. Example results

[1] Results available here: http://archive.nlm.nih.gov/cervigram/research/location/result.htm