Is there a need for biomedical CBIR systems in clinical practice? Outcomes from a usability study

Sameer Antani, Zhiyun Xue, L. Rodney Long, Deborah Bennett, Sarah Ward, George R. Thoma
National Library of Medicine, NIH, Bethesda, MD

ABSTRACT

Articles in the literature routinely describe advances in Content Based Image Retrieval (CBIR) and its potential for improving clinical practice, biomedical research and education. Several systems have been developed to address particular needs, however, surprisingly few are found to be in routine practical use. Our collaboration with the National Cancer Institute (NCI) has identified a need to develop tools to annotate and search a collection of over 100,000 cervigrams and related, anonymized patient data. One such tool developed for a projected need for retrieving similar patient images is the prototype CBIR system, called CervigramFinder, which retrieves images based on the visual similarity of particular regions on the cervix. In this article we report the outcomes from a usability study conducted at a primary meeting of practicing experts. We used the study to not only evaluate the system for software errors and ease of use, but also to explore its "user readiness", and to identify obstacles that hamper practical use of such systems, in general. Overall, the participants in the study found the technology interesting and bearing great potential; however, several challenges need to be addressed before the technology can be adopted.

Keywords: content-based image retrieval, usability study, medical image analysis system, uterine cervix image

1. INTRODUCTION

Researchers in biomedical computer-based image retrieval (CBIR) have long sought the goal of developing systems that can be adopted in clinical practice, or used in medical research or education. Several papers have underlined the significance of content-based indexing for the effective management of visual medical data, have addressed the possible barriers to the use of CBIR in medicine and have identified a set of gaps/limitations in existing CBIR applications [1]. These gaps can be broadly classified into four categories: content gap, performance gap, usability gap, and feature gap [2]. Bridging these gaps could make a CBIR system better positioned for use in the biomedical community. However, is it sufficient to develop technology to address a perceived need for it, for that technology to be accepted and used as expected? Lack of adoption of the technology into any of the target areas and the high number of articles on the topic in the literature compels one to explore the "user readiness" of the technlogy and to identify obstacles that hamper its practical use.

The National Library of Medicine (NLM) has developed a CBIR system, called CervigramFinder, that retrieves images of the uterine cervix based on visually similar characteristics of particular anatomical or pathological regions on the cervix. The image collection and related anonymized patient data were acquired as a part of a multi-year longitudinal study conducted by the National Cancer Institute (NCI) in the continental United States and Costa Rica. Capabilities of the CervigramFinder and related prior work have been reported earlier [3,4,5]. Background about the pathology is briefly covered in Section 2.

NLM conducted a usability test at the 2010 bi-annual conference of American Society for Colposcopy and Cervical Pathology (ASCCP) to evaluate CervigramFinder for ease of use, and to explore the role of such a tool in routine workflow of clinicians and health care providers. The attendees of the meeting are experts in the field of colposcopy and cervical pathology. They are practicing gynecologists, colposcopists, and nurse-practitioners. In this paper we describe the outcomes of the usability test and suggest how to promote adoption of CervigramFinder and similar CBIR systems beyond research and education and into clinical practice.

2. BACKGROUND

Cervical cancer, one of the leading cancers affecting women worldwide, usually develops from persistent infection caused by a certain type of human papillomavirus (HPV) that may be transmitted through sexual contact [6]. It typically takes a number of years to progress from a pre-cancer stage known as cervical intra-epithelial neoplasia (CIN) to a potentially life threatening invasive cancer. Early detection via screening tests is a key to the successful prevention and the treatment of cervical cancer. Among screening techniques, cervicography is routinely used as a low-cost visual screening method. It is an important medical test because a large percentage of new cases of invasive cervical cancer occur in developing countries or poor areas. Cervicography consists of the use of a specialized optical camera to take images of the cervix after exposure to acetic acid solution. These color photographic images of the cervix, referred to as cervicographic images or cervigrams, are then interpreted by gynecologists or trained practitioners who determine whether an *acetowhite lesion* is present and, if so, the degree of abnormality in the cervical tissue. Patients are biopsied for further diagnosis, if necessary. Therefore, cervigrams play a significant role not only in the objective documentation of patient records and clinical findings, but also in the patient diagnosis, surgical planning, and medical treatment of cervical cancer.

Two large studies of cervical cancer which have used cervicography are important for our work: the ALTS Project and the Guanacaste Project, both carried out by the National Cancer Institute (NCI). The ALTS (ASCUS/LSIL Triage Study) project [7] was a multicenter, randomized clinical trial to study the management strategy alternatives for low-grade cervical cytologic abnormalities. It was carried out at four university medical centers in the United States. The Guanacaste project [8] was a prospective cohort study of HPV infection and cervical neoplasia. It was conducted in the Guanacaste province of Costa Rica, an area with a very high prevalence rate of cervical cancer. During these two studies, patients were screened using cervicography and also underwent cytological, colposcopic, and molecular tests after a personal interview. The information collected during both projects includes not only text data on the patients but also associated image data obtained during the screening tests, such as cervigrams. The National Library of Medicine (NLM), in collaboration with the NCI, has managed and maintained the data. NLM has also been developing a set of related searching and analysis software tools to support important clinical studies and training of cervical cancer [9]. One of those tools is the CervigramFinder, a system for searching similar images in the NLM cervical cancer database by visual attributes. It has been developed to expand the capability of another NLM tool, the *Multimedia Database Tool* (MDT), which is a text-only database searching tool for the ALTS and Guanacaste projects. In [3,4], we have reported the system architecture, the features used, and the retrieval performance of the CervigramFinder .

3. UPDATES TO CERVIGRAM FINDER

3.1 Region Segmentation and Labeling

CervigramFinder is a region-based, image-searching tool. Its design is motivated by the observation that physicians look for visual characteristics *within particular regions* when examining the cervix surface to identify abnormality and decide if a biopsy is necessary. In addition, in the cervigrams, only the area in the cervigram image within the cervix anatomical boundary is of significance for our purposes. The unrelated objects outside the cervix, such as vaginal walls, medical instruments, and text labeling, occupy a considerable portion of the image and adversely affect the retrieval performance if included into the analysis. Consequently, region extraction becomes a crucial component of the system. There are two aspects for identifying regions in CervigramFinder: 1) how to create a query region in the query image, and 2) how to extract regions in each image of the entire database. We employed alternative approaches, hoping to combine the knowledge or understanding of physicians/end-users with the computer's computational efficiency.

3.2 Query region

For query region formulation, CervigramFinder implements a user-in-the-loop approach. It provides two options for creating a query region given a cervigram. One is to provide users the list of regions which are clinically important, and which were manually drawn by medical experts, and ask users to select the query region from that list. The other is to let users delineate by hand the boundary of the region they are interested in. The query formulation provided by CervigramFinder tries to alleviate the gap between the user's understanding and the information that

can be extracted by the machine. After the query region is identified, the visual attributes of the region are then extracted by automatic processing algorithms. The extracted feature vector of the query region is then compared with pre-computed feature vectors of all the regions in the database to find the most similar ones.

3.3 Regions in the database

For regions in the database, three tools have been developed for segmenting/labeling clinically-significant regions. They are: 1) the *Boundary Marking Tool* (BMT); 2) the *Cervigram Segmentation Tool* (CST); and 3) the *Multi-Observer Segmentation Evaluation System* (MOSES). These tools demonstrate our work not only in proposing new algorithms but also in developing practical tools. They also illustrate our vision in attempting to unify three crucial topics in medical image segmentation: ground truth collection, automatic image segmentation, and segmentation evaluation.

The BMT (Figure 1) is for collecting "ground truth data" [10]. It allows medical experts to manually draw the boundaries of regions and to enter the labels of the regions as well as detailed interpretative information relevant to the evaluation of the cervix. Data marked and collected using the BMT has been used in multiple NCI studies, resulting in several medical publications, by both gynecological experts and image processing researchers in cervical cancer. One use of this tool was to mark two types of important regions, cervix boundary and acetowhite lesions, in a dataset of 939 cervigrams, by twenty clinicians with varying numbers of years of experience. This resulted in a set of manual markings for each of the 939 cervigrams; because of the study design, and variability in expert opinion, the number of markings on each cervigram was not constant, and in fact ranged from one to twenty.

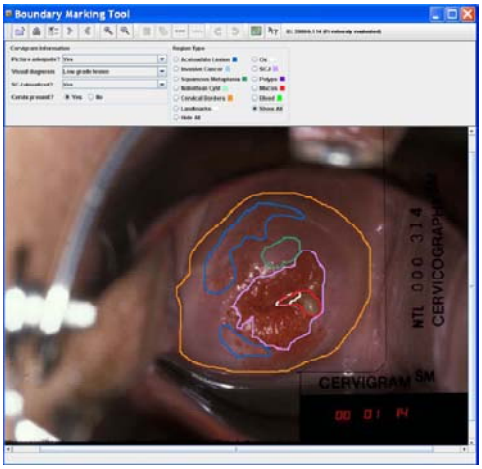

Figure 1. Screen captures of the BMT's interface

In addition to BMT, we have been working on developing algorithms for automatic segmentation and labeling of important regions in the cervigrams, since the time required for manual segmentation would be prohibitive for such a large archive. Automatic analysis of cervigrams is a very difficult task due to the large diversity of the images in the database and the intrinsic content complexity of these images. We have developed a multi-step segmentation approach to address this difficulty. To promote collaboration between engineers and physicians and to receive feedback on our cervigram segmentation algorithms, we have been developing the CST (Figure 2) [10], which includes the steps of cervix region-of-interest coarse detection, specular reflection elimination, os detection, and columnar epithelium segmentation.

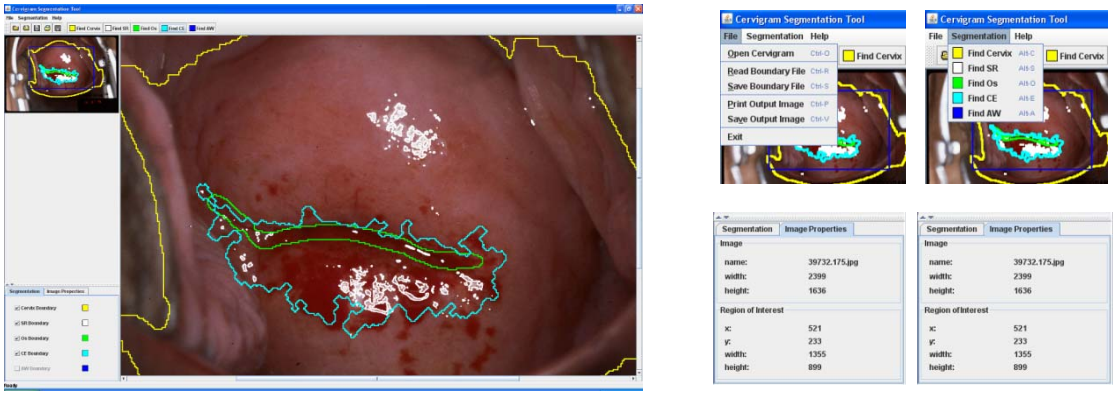

Figure 2. Screen captures of the user interface of the CST

Finally, the tool MOSES (Figure 3) [11] has been developed, with the goals of 1) quantitatively analyzing the inter- and intra-expert variability; 2) objectively evaluating the effectiveness of automated segmentation; and 3) generating a reliable ground truth data with higher consensus for cervigram segmentations. It uses Bayesian Decision Theory and the MAP optimization principle to evaluate multiple observers' segmentations, and handles more scenarios than the well-known STAPLE algorithm.

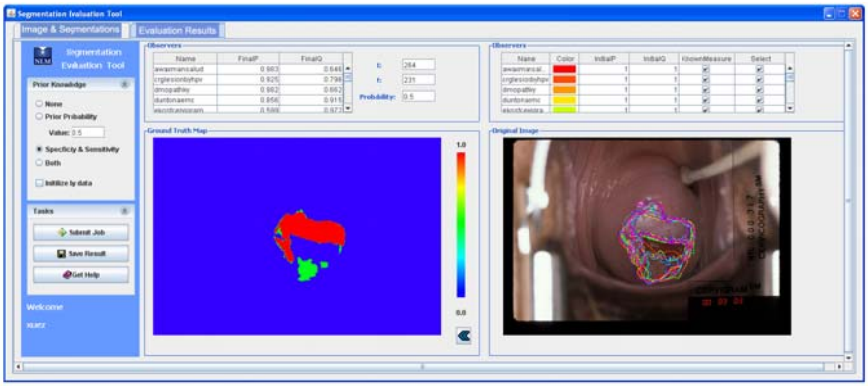

Figure 3. Screen captures of the user interface of the MOSES

3.4 User Interface

Compared to the version of CervigramFinder reported in [3,4], several changes to the user interface have been made. Major modifications include: modification to achieve consistent look-and-feel across Windows and Mac OS; addition of a monitor calibration page; addition of user "login/logout" functionality; addition of the function of saving the query and corresponding retrieved results to our server; updating the help files; and addition of the information/instruction box. We have also added a very detailed patient record (extracted from the MDT system) to provide a better understanding of the visual features of each cervigram. The patient record includes patient age, the cervigram visual examination result, the colposcopic examination result, the HPV infection types, the referral diagnosis, and the final incident diagnosis. We believe this is a step toward providing greater context to image information – thereby providing some semantics. We consider this to be only a first step, and we plan greater correlation between patient record data and image features for the future.

Figure 4. User interface of CervigramFinder

4. USABILITY STUDY: METHOD, OUTCOMES AND DISCUSSION

Two of our longstanding goals are to bridge the disconnect between systems developed for research and those that can be effectively used by clinicians working in the field, in this case colposcopy and cervical gynecology; and to find other potential uses for our software tools. Toward these goals, we took advantage of a gathering of experts at the biennial conference of the American Society of Colposcopy and Cervical Pathology (ASCCP), which is the national organization in the field. It aims to provide a forum for education and dissemination of knowledge about the lower genital tract through the use of colposcopy, and includes the disciplines of pathology, cytology, cytogenetics, preventive medicine, basic research, gynecologic oncology, and endocrinology, which are relevant to the understanding of the patho-physiology of the disease processes of the lower genital tract.

Method: We developed and administered both a paper-based survey and a hands-on, software usability study to create a baseline evaluation of CervigramFinder. The survey was designed to help determine the level of interest of specialized physicians and healthcare practitioners in the field in accessing the NLM tools: BMT, and CervigramFinder, as well as to determine the points of interests most relevant to work conducted by these specialists.

The usability study was conducted during an hour-long session in a "think out loud" format where the participants were encouraged to voice their thoughts on the tool as they used it, while an expert facilitator guided the progress of the session. The session was divided into two parts. In the first part, the users evaluated the BMT and marked images using the tool. Test scenarios were described to the participants that related to patient care or medical research. In the second part, the participants evaluated CervigramFinder with a similar setup, where the scenarios were follow-on from the BMT session. Participants were first asked a number of open-ended questions about their past experience working with digital images of the cervix. They were then asked to "think out-loud" while performing a set of questions regarding the CervigramFinder Web site, followed by a set of two tasks using that application. Finally, the participants were asked a series of follow-up questions about their experience using CervigramFinder and were asked to complete a second usability questionnaire related to the application.

Outcome and discussion: One hundred and two (102) attendees of the ASCCP meeting completed the paper-based survey on-site. Of the survey participants, ninety five percent (95%) were clinicians directly involved in patient care; the rest were educators and/or researchers. Of those surveyed, sixty three percent (63%) maintained records for patients using paper notes with marks on a diagram of the cervix, while only nine percent (9%) utilized some form of an electronic record with markup on sketched drawings of a cervix. None of the participants maintained images of the patient's cervix in routine clinical use. A small fraction indicated that they sometimes used images acquired

from a colposcope for patient education. The three features related to computer software that respondents felt were most important to their work were:

1. It should have a simple interface,
2. It should have a fast response to their actions, and
3. It should let them view medical images (preferably for all patient visits, and at least for those acquired in a prior visit, for purposes of comparison and study of disease or treatment progress.)

Fifty-eight percent (58%) of the respondents felt it was very important that the software allow them to annotate medical images, while forty one percent (41%) felt that it was important for them to be able to exchange case data with their colleagues. This is a compelling argument for a common transport format. This is probably not new to those practicing radiology and with expertise in DICOM; but it is novel to the field of colposcopy. Fifty two (52%) of respondents wanted to use the CervigramFinder for biomedical education of their students. This suggests that there is a need for systems with image retrieval capabilities and atlas-like capabilities. However, due to the lack of sufficient use of electronic medical records (EMR) and image management, it is unclear how such systems will have the desired acceptance in routine use.

Twelve (12) clinicians were recruited as participants in the usability study from registered attendees of the ASCCP meeting. We measured the effectiveness of the CervigramFinder application by tallying the number of tasks each participant completed without difficulty. Participants were assigned a score ranging from 0 to 2 based on their ability to perform the study tasks. We gauged the efficiency of the tools by measuring the time it took for the participants to complete each task scenario. We determined user impressions of the application by asking them to rate the tools on ease of use both before and after using the site. In addition, at the end of the evaluation, we administered a questionnaire based on the System Usability Scale (SUS) [12] to the participants to measure their satisfaction with the applications. We recorded participant reactions to the Web site used to access the applications by compiling their comments, observations, and suggestions for improvement.

Our test participants uncovered a number of usability issues while evaluating the CervigramFinder. One of the most significant findings was the lack of basic computer skills in the health care professionals participating in the study. The typical user had trouble browsing for files and understanding icons and terminology used in the tool's interface. Their previous experience appeared to be limited to using specialized medical systems and explicit workflows within these systems. They also had difficulty distinguishing between the features of the tools and the Web browser they were using. (CervigramFinder is designed as an applet and runs within a Web browser.)

Participants overwhelmingly felt that the most useful potential application of the CervigramFinder would be as a teaching tool and were very interested in an application that could provide them with a library of images that includes good examples of multiple types of diagnoses. They also felt that the tool could be quite useful in conducting quality control assessments within their own practices. Participants liked the idea of searching based on image features; however, they expressed interest in also searching by diagnosis or condition. Users also uncovered many problems with the tool. They reported that it was challenging to obtain a clear understanding of the purpose and functionality of the tool without any training on the tool's capabilities. They could not discern how to properly formulate a visual query, which is a critical component in CBIR systems, and nearly all users failed to use the drawing tools successfully, reiterating a performance gap identified earlier. They also had other user interface concerns that hindered their use of the system.

Participants gave the CervigramFinder an average score of 53.86 on the System Usability Scale [12], indicating that they disliked the system in its current state. It is important to contrast this against their limited computing capability and lack of experience with EMR and routine use of archived images. However, in a positive light the score suggests that there is a need for a well-designed biomedical CBIR system that works well with EMR systems. The outcomes of the results are being used to develop a new design for the CervigramFinder user interface.

5. CONCLUSIONS AND FUTURE WORK

The inclusion of visual features in medical image search systems to retrieve cases is a very important research topic. For searching and retrieving images from a large archive of cervigram database collected through two important cervical cancer studies, we have developed a prototype CBIR system, CervigramFinder. Unlike many CBIR

systems, CervigramFinder provides the capability for visual search of clinically important regions using *local image characteristics* that are significant for cervigrams. Initial experiments with CervigramFinder on a subset of the NLM image repository have demonstrated the potential of this system. However, acceptance and adoption of CervigramFinder, or any other CBIR system, may depend heavily on close interaction with the medical practitioners, researchers, and/or educators who are the end users. Further, such technology, it appears, may face obstacles as a standalone system. It may have greater success where it is available as a feature in other patient record systems in routine clinical use. The study also highlighted various gaps that still persist and must be minimized or overcome before it is accepted. As an immediate step, we have improved CervigramFinder with the integration of associated patient records. Other recommendations from the study will be incorporated in future revisions of the system. It is a goal for us to develop CervigramFinder, or other such systems, into tools that are a valuable part of routine clinical care and a complementary component to current text-based tools.

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. S. Antani, L. R. Long, G. R. Thoma, “Bridging the gap: Enabling CBIR in medical applications,” Proc. 21st IEEE International Symposium on Computer-Based Medical Systems, pp. 4-6, Jyväskylä, Finland. June 2008.
2. T.M. Deserno, S. Antani, R. Long, “Ontology of Gaps In Content-Based Image Retrieval.” Journal of Digital Imaging. April 2009;22(2):202-15.
3. Z. Xue, S. Antani, L. R. Long, J. Jeronimo, G. Thoma, “A Web-accessible content-based cervicographic image retrieval system”, Proc. of SPIE Medical Imaging, Vol. 6919, pp. 691907-1-9, February, 2008.
4. Z. Xue, L. R. Long, S. Antani, G. R. Thoma, “A system for searching uterine cervix images by visual attributes”, Proceedings of the 22nd IEEE International Symposium on Computer Based Medical Systems, August, 2009
5. Z. Xue, S. Antani, L. R. Long, G.R. Thoma, “Web-accessible cervigram automatic segmentation tool”, Proceedings of SPIE Medical Imaging, Volume 7628, pp. 76280Z-76280Z-8, 2010.
6. J.W. Sellors and R. Sankaranarayanan, “Colposcopy and treatment of cervical intraepithelial neoplasia - A beginner’s manual”, Edited by J.W. Sellors and R. Sankaranarayanan, Published by the International Agency for Research on Cancer, France, 2003.
7. M. Schiffman, M. E. Adrianza, “ASCUS-LSIL triage study: design, methods and characteristics of trial participants”, Acta Cytol, Vol. 44, No. 5, pp. 726-742, 2000.
8. 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.
9. http://archive.nlm.nih.gov/proj/impactOfTools.html, last accessed on December 30th, 2010.
10. 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, Vol.10, No. 1, pp.16-25, January 2006.
11. Y. Zhu, W. Wang, X. Huang, D. Lopresti, L. R. Long, S. Antani, Z. Xue, G. R. Thoma, “Balancing the role of priors in multi-observer segmentation evaluation”, Journal of Signal Processing Systems. New York, NY. May 2008.
12. J. Brooke, "SUS: a ‘quick and dirty’ usability scale". in P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland. Usability Evaluation in Industry. London: Taylor and Francis. 1996.