MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Content-Based Image Retrieval for Large Biomedical Image Archives

Sameer Antani, L. Rodney Long, George R. Thoma

*Lister Hill National Center for Biomedical Communications, National Library of Medicine,
National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA*

Abstract

Content-Based Image Retrieval (CBIR) has been a topic of research interest for nearly a decade. Approaches to date use image features for describing content. A survey of the literature shows that progress has been limited to prototype systems that make gross assumptions and approximations. Additionally, research attention has been largely focused on stock image collections. Advances in medical imaging have led to growth in large image collections. At the Lister Hill National Center for Biomedical Communication, an R&D division of the National Library of Medicine, we are conducting research on CBIR for biomedical images. We maintain an archive of over 17,000 digitized x-rays of the cervical and lumbar spine from the second National Health and Nutrition Examination Survey (NHANES II). In addition, we are developing an archive of a large number of digitized 35mm color slides of the uterine cervix. Our research focuses on developing techniques for hybrid text/image query-retrieval from the survey text and image data. In this paper we present the challenges in developing CBIR of biomedical images and results from our research efforts.

Keywords:

Image Processing, Information Storage and Retrieval, Multimedia Databases, Medical Informatics Applications

Introduction

Content-based image retrieval (CBIR) has attracted much research interest in recent years [1]. In particular, there has been growing interest in indexing biomedical images by content [2, 3, 4, 5]. Manual indexing of images for content-based retrieval is cumbersome, error prone, and prohibitively expensive. Due to the lack of effective automated methods, however, biomedical images are typically annotated manually and retrieved using a text keyword-based search. A common drawback of such systems is that the annotations are imprecise with reference to image feature locations, and text is often insufficient in enabling efficient image retrieval. Even such retrieval is impossible for collections of images that have not been annotated or indexed. Additionally, the retrieval of interesting cases, especially for medical education or building atlases, is a cumbersome task. CBIR methods developed specifically for biomedical images could offer a solution to such problems, thereby augmenting the clinical, research, and educational aspects of biomedicine. For any class of biomedical images, however, it would be necessary to develop suitable feature representation and similarity algorithms that capture the "content" in the image.

The Lister Hill National Center for Biomedical Communications, a research and development division of the U.S. National Library of Medicine (NLM), maintains a digital archive of 17,000 cervical and lumbar spine images collected in the second National Health and Nutrition Examination Survey (NHANES II) conducted by the National Center for Health Statistics (NCHS) (figure 1 (a) and (b)). Classification of the spine x-ray images for the osteoarthritis research community has been a long-standing goal of researchers at the NLM, and collaborators at NCHS and the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS). Also, the capability to retrieve these images based on geometric characteristics of the vertebral structures is of interest to the vertebral morphometry community. Medical experts have identified visual features of the images specifically related to osteoarthritis, but the images have never been manually indexed for these features which include anterior osteophytes, disc space narrowing for the cervical and lumbar spine, subluxation for the cervical spine, and spondylolisthesis for the lumbar spine. Another archive of 100,000 digitized 35mm color slides of the uterine cervix is being created in collaboration with the National Cancer Institute (NCI) (figure 1(c)). Researchers at NCI would like to enable use of these images for research and training at sites around the world. The design of a system to achieve these ends relies on research in image compression, database management, and CBIR for image query on the uterine cervix images.

Automated or computer-assisted classification, query, and retrieval methods for large medical image archives are highly desirable, since such methods offset the high cost of manual classification and manipulation by medical experts. We are investigating automated or computer-assisted methods that use image features for indexing and retrieval of these images in a manner acceptable to the biomedical community. In addition, we are devoting research efforts into classification of pathology, such as the detection of presence of anterior osteophytes, disc space narrowing, subluxation, and spondylolisthesis in spine images; and squamo-columnar junction boundary, regions with acetowhitening, vasculature, mosaicism and punctation, on the uterine cervix images.

As an initial step, we have implemented a modular prototype CBIR system for a subset of the spine x-rays and the associated health survey text data [6]. The system supports retrieval based on shape similarity to a sketch of a complete or partial vertebra, an example vertebral image, as well as conventional text retrieval. In this paper we present the technical considerations in developing a system for CBIR of medical images, open research problems, and the lessons learned from our research efforts.

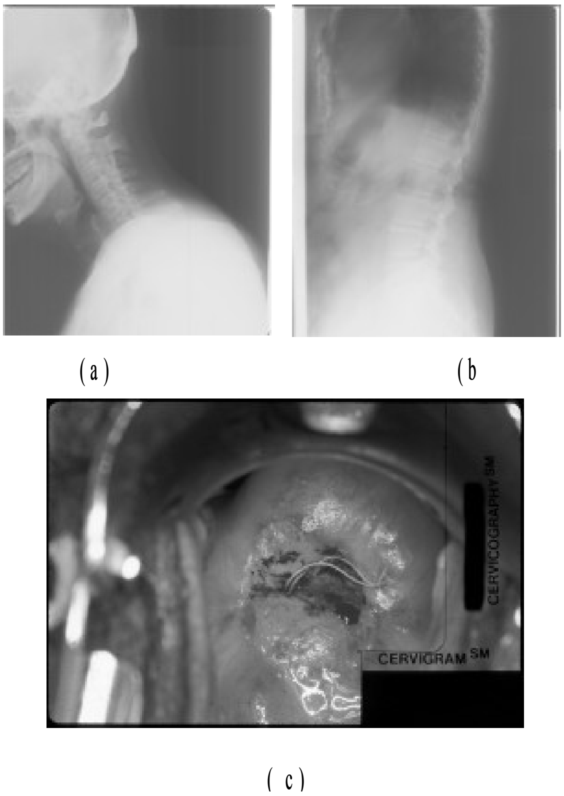


Figure 1 - Example images of (a) cervical and (b) lumbar spine x-ray from NHANES II collection, and (c) uterine cervix.

The CBIR Trail

Content-based image retrieval hinges on the ability of the algorithms to extract pertinent image features and organize them in a way that represents the image content. Additionally, the algorithms should be able to quantify the similarity between the query visual and the database candidate for the image content as perceived by the viewer. Thus, there is a systemic component to CBIR and a more challenging semantic component. As a first step, the images must be indexed, at least, for the pathologies of interest.

Indexing Trail

The indexing trail (figure 2) has been presented from a systemic viewpoint. A graphic user interface (GUI) of the indexing system allows the users to index the text and image data. In indexing images, visual features that correspond to the pathology of interest are segmented (extracted) from the image. Shown as the "Segmentation" block in the figure, this step is synonymous with "Feature Extraction". The output of the segmentation step is usually in the form of image components such as subimages, edges, boundary contours, color/intensity measurements, texture measurements, etc. Feature extraction is usually done at the local region of interest. In case of the digitized spine x-ray image collection, the only features of interest are the shape of the vertebra and the positional relationship with other vertebrae. At the end of the segmentation step the resulting data is a vector of 2D coordinate points describing the vertebra boundary outline. Segmentation techniques include variants of active contour segmentation [7] and active shape modeling [8]. In the case of the uterine cervix images, the extracted features include, in addition to shape, color and texture measurements on homogeneous regions that are determined using a *k*-Nearest Neighbor classifier. Examples of extracted features overlaid on images are shown in figures 3 (a) and (b).

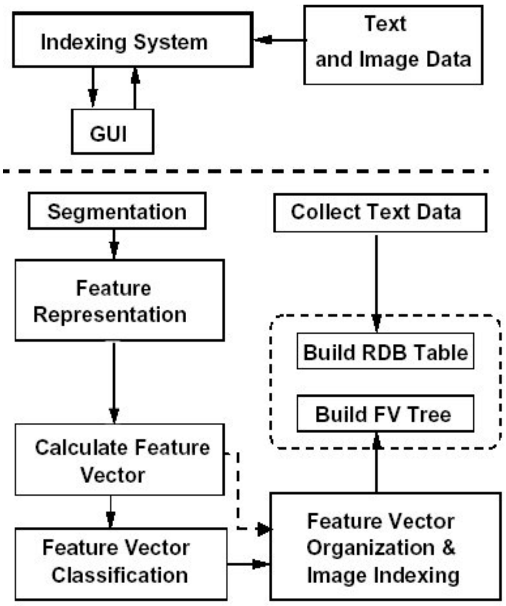


Figure 2 - CBIR Indexing Trail

The segmented features, then, need to be represented in a form suitable for indexing and similarity computation. This task is done at the "Feature Representation" stage. "Feature Vector Computation" is often coupled with this stage. The output of these stages is a feature vector that is often in a form invariant to rotation, translation, scale, and also possesses many other properties such as *stability, uniqueness, etc* [6]. Image similarity is, then, defined as the distance between the feature vectors for two images. Also, each feature representation algorithm may have to use a corresponding similarity measure. A side effect of feature representation is loss of information incurred due to approximation which is done for algorithmic or efficiency reasons, or to avoid the *curse of dimensionality.* Representations, such as histograms or color averages, approximated boundaries, are often sufficient to enable some form of CBIR and are found in many prototype systems discussed in the literature [1]. The cost in loss of representation of subtle variations in image features, however, can lead to poor retrieval quality. Storing higher dimension feature vectors, while enabling query of subtleties in image content, can cause problems for indexing, creating a Catch-22 situation. We are experimenting with a variety of shape representation techniques for segmented vertebrae that include polygon approximation, Fourier descriptors, shape properties, invariant moments, and Procrustes distance [6, 9, 10]. An outstanding problem in the extraction of feature vectors from the raw boundary data is development of an effective shape representation and similarity method that provides for data reduction while simultaneously preserving the shape characteristics that are essential for reliable indexing and retrieval.

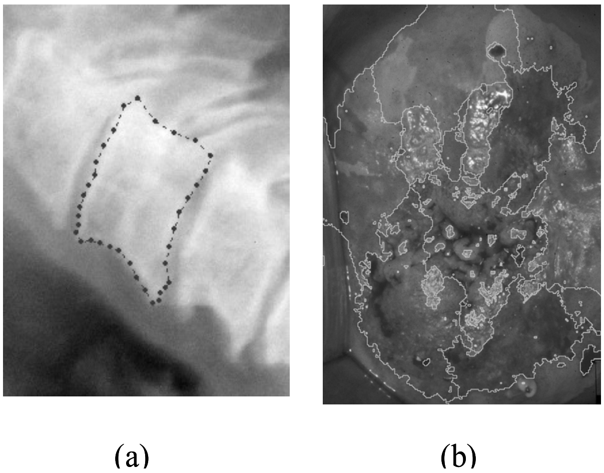


Figure 3 - Examples of segmented features: (a) c-spine vertebra boundary; (b) regions on the uterine cervix

.

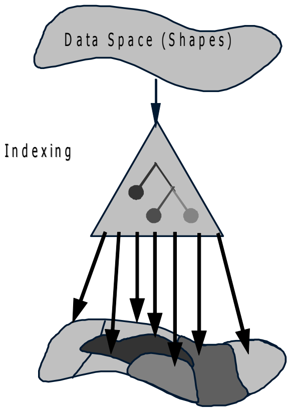


Figure 4 - Hierarchical cluster indexing tree

Image content represented by feature vectors is then indexed. Unlike in a traditional database, however, it is difficult to develop unique keys from the feature vectors. One approach is to use hierarchical cluster trees (figure 4). This approach links images to leaf nodes in a tree. It then clusters "similar" images together and assigns their cluster centroids as parent nodes. This process is repeated on these centroids until only one remains. Such a hierarchical organization strategy can be very efficient and significantly reduces image search times. In addition, it supports both *target queries*, where one matching image is sought, as well as *range queries*, where images that match a certain feature measurement range are sought. There are, however, some shortcomings with this approach. First, it requires that the similarity measure be a metric and most effective similarity measures are relative. Second, the index tree is optimized for a single type of query, e.g., in spine x-ray images, the tree might be optimized for queries on anterior osteophytes. For other query types a new index tree would be necessary. This limits the types of queries possible on a dataset and is not directly helpful to the long term goals of CBIR. As an initial step, however, we have adopted this approach [11] for organizing indexing trees and optimizing the node structure with the spine x-ray images shapes. In general, organization of image features for CBIR is an open research problem. The survey text data that accompanies the images is indexed in a traditional RDBMS. Our current implementations do not link images indexed in the hierarchical cluster tree to the RDBMS text data, though such an approach is conceivable. Currently, we link the image to the text data by the image name.

Retrieval Trail

The retrieval trail (figure 5) repeats many of the initial steps from the indexing trail. Two types of queries have been discussed in the CBIR literature, query by (image) example and query by feature. In addition, one may augment these visual queries with text fields. These hybrid text/image queries can be very useful in medical image databases which often have supporting text information for the image data. Hybrid queries can significantly reduce the search space and improve retrieval efficiency. Additionally, interesting variations in combining text and image query results can be developed to further enhance system utility

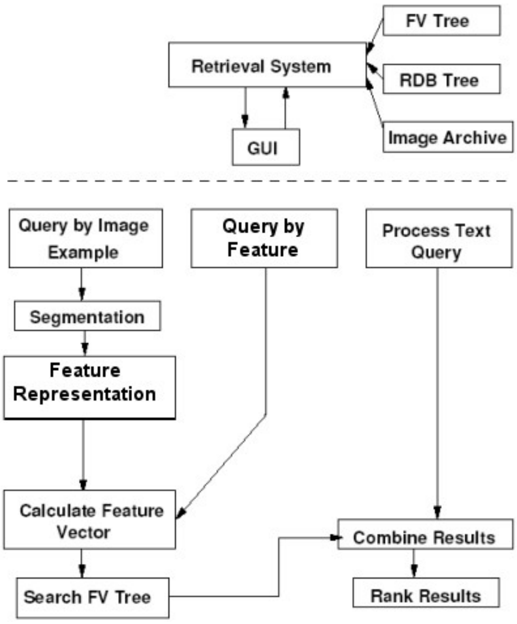
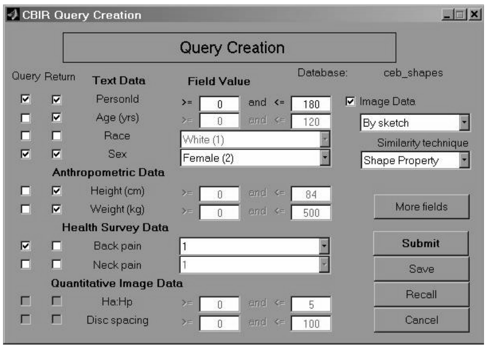


Figure 5 - CBIR Retrieval Trail

The retrieval system uses the feature vector (FV) tree, the text RDBMS, and the image archive as inputs. The GUI permits the user to query by image example or by feature. Additionally, for hybrid queries, the user may modify any of the text fields. Figures 6-8 show screen shots for the queries made to the prototype CBIR software developed in MATLAB for spine x-ray images. In case of query by image example, the user marks the region of interest on a query image (figure 7). The retrieval system extracts necessary features and represents them in the same form as that for the archived images. For query by feature, the feature is specified by the user. For example, in case of spine x-ray images, the user draws a sketch of desired vertebra shape (figure 8). It is conceivable that for color images one could specify a desired color and texture. A well designed GUI could assist the user in forming a query. The FV tree is, then, searched for images with similar features and the results are combined with those returned from the text RDBMS (if any) and ranked in similarity distance order and presented to the user (figure 9).


CBIR Query Creation

Query Creation

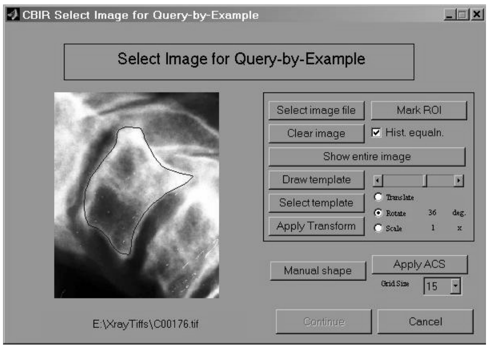
Database: ceb_shapes

Query	Return	Text Data	Field Value			
☑	☑	PersonId	>=	0	and <=	180
□	☑	Age (yrs)	>=	0	and <=	120
□	□	Race	White (1)			
☑	☑	Sex	Female (2)			
		Anthropometric Data				
□	☑	Height (cm)	>=	0	and <=	84
□	☑	Weight (kg)	>=	0	and <=	500
		Health Survey Data				
☑	□	Back pain	1			
□	□	Neck pain	1			
		Quantitative Image Data				
□	□	Ha:Hp	>=	0	and <=	5
□	□	Disc spacing	>=	0	and <=	100

☑ Image Data: By sketch

Similarity technique: Shape Property

Figure 6 - Text query screen shot


CBIR Select Image for Query-by-Example

Select Image for Query-by-Example

☑ Hist. equaln.

○ Translate
● Rotate 36 deg.
○ Scale 1 x

Grid Size: 15

E:\XrayTiffs\C00176.tif

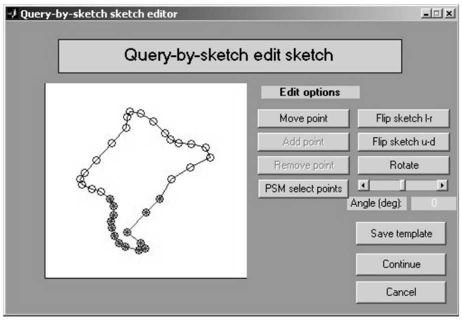
Figure 7 - Query by image example screen shot

Lessons Learned

Techniques covered in CBIR literature for stock images do not meet retrieval reliability necessary for biomedical images. In this section, we present challenges, open problems, and lessons learned for each task in the CBIR trail during our development and use of the research prototype system.

Segmentation: Fully automated segmentation/feature extraction from medical images is a very challenging problem with no method directly applicable as found in the literature. Its quality is affected by three important factors; viz., technique, image quality, and image size/resolution. Hence, computer assisted segmentation is usually a more promising approach. We have faced significant problems with poor image quality in the digitized spine x-ray images where segmentation methods often confuse tissue and vertebra boundaries. We discovered that techniques that have been known to work well for smaller images often do not scale well when presented with higher resolution images. Poor segmentation results were observed on application of fairly robust color and texture segmentation techniques developed for 192x128 pixel Blobworld [12] images to the 2399x1637 pixel uterine cervix images from the NCI dataset.

Feature Representation: The dimensionality of the represented feature is highly correlated with the quality and inversely with efficiency of retrieval. Additionally, developers of a CBIR system must ensure that the representation technique faithfully captures the image content/semantics. We learned this lesson during performance evaluation on retrieval quality of vertebrae with anterior osteophytes. The method selected for representing vertebra boundaries was Fourier descriptors following initial reduction in dimension using polygon approximation. It was found that while the method did fairly well (70% retrieval precision [13]) on complete shape queries, the users were unable to focus on specific regions of the pathology. Intuitive modifications to the query shape were not comparably represented resulting in low quality retrieval. This led us to pursue research into *partial* or *incomplete* shape matching (figures 8 and 9).

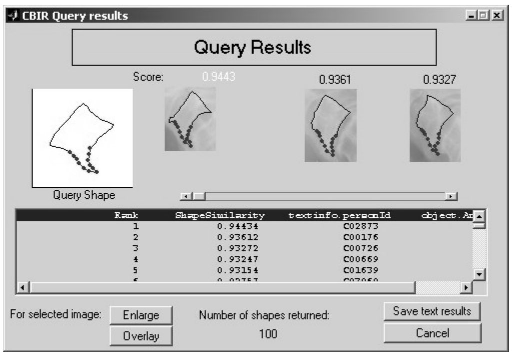

Query-by-sketch sketch editor

Query-by-sketch edit sketch

Angle (deg): 0

Figure 8 - Query by feature (sketch/shape) screen shot: Partial shape selection shown as dark segment.

In a recent evaluation of retrieval performance improvement through the use of partial shape matching, it was observed that for pathological shapes the retrieval relevance improved by a factor of 2.2. That is, the average retrieval performance improved from 3.815 relevant matches with a standard deviation of 1.66 in the top 10 retrieved shapes to on average 8.125 relevant matches with a standard deviation of 0.25[14].


CBIR Query results

Query Results

Score: 0.9443 0.9361 0.9327

Query Shape

Rank	ShapeSimilarity	textinfo.personId	object.Ar
1	0.94434	C02873	
2	0.93612	C00176	
3	0.93272	C00726	
4	0.93247	C00669	
5	0.93154	C01639	

For selected image:

Number of shapes returned: 100

Figure 9 - CBIR results for a partial shape query. Query shape with selected shape segment is shown on left.

Feature organization: Our initial evaluation of hierarchical clustering showed that there is a significant performance improvement over linear searches. In a set of experiments on a 1298 shape database, on average only 176 leaf nodes needed to be accessed for a 13 nearest neighbor (most similar shapes) query. In scaling the range to 40 nearest neighbors, it needed to access an average of only 315.25 nodes. This is a significant improvement over 1298 accesses required in a linear search. It is, however, optimized for only one feature. Additionally, it uses shape descriptions with a fixed number of boundary points since it requires the similarity measure to be a metric. The Procrustes distance [10] that we use as a similarity measure is a metric, but operates only on boundaries with fixed number of points. Fixed number of points in low dimension can affect image feature detail and consequently adversely affect retrieval quality.

GUI Design: CBIR queries are inherently visual in nature. The GUI should allow the users the flexibility to make a variety of queries, create relationships between image features, assign importance to features, and intelligently combine text and image queries to form hybrid queries.

Open Problems

While we have developed reasonably functional solutions for segmentation and representation of vertebrae, these remain open problems for the uterine cervix images. Organization of feature vectors comprising multiple features also remains an open problem. Specifically, it is necessary to decrease the dependence of hierarchical cluster trees on similarity distance metric. Finally, an important problem little discussed in the literature but of much importance, is that of validation of retrieval results. For example, how can we justify calling one set of shape retrieval results better than another? How can we compare results among different shape representations and similarity measures? The validation of the query results in either a quantitative sense or with a non-quantitative approach that will justify confidence in the results using a particular method remains a critical issue for this work.

Beyond the important issue of what we have called "engineering validation" of results, there remains the further issue of biomedical validation, for the biomedical community is the system end user. There is a critical need for more extensive expert data to enable development of better algorithms. A key requirement of these data sets is that they should be collected by multiple expert observers; only then can the performance of computerized methods relative to human performance be evaluated.

With the NCI uterine cervix image set, an open problem is the technique for combining multiple image features. The images are rich in color, exhibit texture in pathology and also have boundaries. Techniques for effective and efficient combination of features need to be developed.

Conclusions

This paper presents our experiences, techniques adopted, and lessons learned in continuing research towards enabling CBIR for large medical image archives at the National Library of Medicine. The paper presents the CBIR trail and presents some results from our work in each task on the trail and open research problems.

References

[1] Antani S, Kasturi R, and Jain R. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition*, 35(4): 945-65, 2002.

[2] Tagare HD, Jaffe CC, Duncan J. Medical image databases: a content-based retrieval approach. *J Am Med Inform Assoc*. 1997 May-Jun; 4(3):184-98.

[3] U. Sinha, A. Ton, A. Yaghmai, R. K. Taira, and H. Kangarloo. Image Content Extraction: Application to MR Images of the Brain. *RadioGraphics*, March 1, 2001; 21(2): 535-547

[4] M. E. Mattie, L. Staib, E. Stratmann, H. D. Tagare, J. Duncan, and P. L. Miller PathMaster: Content-based Cell Image Retrieval Using Automated Feature Extraction. *J. Am. Med. Inform*. Assoc., July 1, 2000; 7(4): 404-415.

[5] Long LR, Antani S, Lee DJ, Krainak DM, Thoma GR. Biomedical information from a national collection of spine x-rays: film to content-based retrieval. *Proc SPIE Med Imaging: PACS and Integrated Med Info Sys: Design and Evaluation*, 15-20 Feb 2003, San Diego, CA, SPIE Vol. 5033: 70-84.

[6] Antani S, Long LR, Thoma GR. A biomedical information system for combined content-based retrieval of spine x-ray images and associated text information. *Proc. 3rd Indian Conf on Computer Vision, Graphics, and Image Proc.*, Ahmedabad, India, 16-18 Dec 2002, 242-47.

[7] Kass M, Witkin A., Terzopoulos, D. Snakes: Active contour models. *Int. J Computer Vision*, 1:321-331, 1987.

[8] Cootes, T.F., Taylor, C.J. Statistical models of appearance for medical image analysis and computer vision. *Proc. SPIE Med Imaging: Image Proc*, 17-23 February 2001, San Diego, CA, SPIE Vol. 4322: 236-48.

[9] Xu, X, Lee D.J, Antani S. et al. Localizing contour points for indexing an x-ray image retrieval system. *Proc. 16th IEEE Symp Computer-based Med Sys,* New York, NY, 26-27 June 2003, 169-74.

[10]Dryden IL, Mardia KV. *Statistical Shape Analysis*. John Wiley & Sons; 1998.

[11]Qian X, Tagare HD. Optimally adapted indexing trees for medical image databases. CDROM *Proc. IEEE Intl. Symp. Biomedical Imaging*, Washington, D.C., 7-10 Jul 2002.

[12]Blobworld. http://elib.cs.berkeley.edu/photos/blobworld.

[13]Antani S, Long LR, Thoma GR, Lee DJ. Anatomical shape representation in spine x-ray images *Proc 3rd IASTED Int Conf Visualization, Imaging and Image Proc* (VIIP 2003), Sept 2003; Vol. 1:510-15.

[14]Antani S, Xu X, Long LR, Thoma GR. Partial Shape Matching for CBIR of Spine X-ray Images. *Proc IS&T/ SPIE Electronic Imaging - Storage and Retrieval Methods and Applications for Multimedia* 2004. Jan 2004; SPIE Vol. 5307: 1-8.

Address for correspondence

Sameer Antani, Ph.D.
Lister Hill National Center for Biomedical Communications,
National Library of Medicine,
8600 Rockville Pike, Mail Stop 55 Bethesda, MD, 20894, USA.
Email: antani@nlm.nih.gov