A Web Collaboration System for Content-Based Image Retrieval of Medical Images

Dave Tahmoush and Hanan Samet
University of Maryland, College Park, Maryland USA

Abstract

Building effective content-based image retrieval (CBIR) systems involves the combination of image creation, storage, security, transmission, analysis, evaluation feature extraction, and feature combination in order to store and retrieve medical images effectively. This requires the involvement of a large community of experts across several fields. We have created a CBIR system called Archimedes which integrates the community together without requiring disclosure of sensitive details. Archimedes' system design enables researchers to upload their feature sets and quickly compare the effectiveness of their methods against other stored feature sets. Additionally, research into the techniques used by radiologists is possible in Archimedes through double-blind radiologist comparisons based on their annotations and feature markups. This research archive contains the essential technologies of secure transmission and storage, textual and feature searches, spatial searches, annotation searching, filtering of result sets, feature creation, and bulk loading of features, while creating a repository and testbed for the community.

1. Introduction

The number of digital medical images is rapidly rising, prompting the need for improved storage and retrieval systems. Image archives and imaging systems are an important economic and clinical factor in the hospital environment [16]. The management and the indexing of these large image repositories is becoming increasingly complex. Most retrievals in these systems are based on the patient identification information or image modality [8] as it is defined in the DICOM standard [12], but it is hoped that inclusion of other features can improve the effectiveness of this type of system. Archimedes includes retrieval based on features and combinations of features, as well as on patient identification information, doctor's notations, and image modality in order to develop effective CBIR. Archimedes also includes filtering of the result set in order to further refine and improve the search.

Clinical decision support techniques such as case-based reasoning [7] or evidence-based medicine [2, 3] rely on effective CBIR development. Image and visual feature-based searches will help find similar images, but textual searches are always going to be an important part of any medical CBIR system, especially

The support of the National Science Foundation under Grants EIA-00-91474 and CCF-0515241, Microsoft Research, and the University of Maryland Graduate Research Board is gratefully acknowledged.

Proceedings of SPIE Medical Imaging 2007 - PACS and Imaging Informatics- Vol.6516 - San Diego, USA, Feb 2007

through searches on patient information or characteristics. That is why searching on patient information and other text is already supported in the Archimedes system.

The integration of CBIR methods into Picture Archiving and Communication Systems (PACS) has been proposed several times. PACS are the main software components used to store and access the large amount of visual data in medical departments. Often, several layer architectures exist for quick short-term access and slow long-term storage [9], but this is becoming increasingly unnecessary as technologies have improved. The Archimedes system was designed as a web-based system for both the development and evaluation of CBIR, and provides a platform to evaluate the usefulness and effectiveness of incorporating CBIR changes into PACS.

Several frameworks for distributed image management solutions have been developed such as I2Cnet [10, 11]. Image retrieval based on visual features is often proposed but unfortunately little is said about the visual features used or the performance obtained. The difficulty is in getting so many talented people in different fields together to work on all of the aspects of a project. A real medical application of CBIR methods and the integration of these tools into medical practice have required a large group in very close cooperation for a long period of time. CBIR systems that have followed this model are the Assert system for the classification of high resolution CTs of the lung [1, 14] and the IRMA system for the classification of images into anatomical areas, modalities and view points [6]. The Archimedes system hopes to bypass this difficulty with a web-based community of researchers who can contribute features, images, results sets, diagnoses, and other expertise in an open research environment. This paper demonstrates a technology to decentralize this process by including a large web-based collaboration of partners, each achieving individual goals while contributing to the overall goal of an improved CBIR system.

Comparing CBIR systems is often challenging because commercial companies are often unable or unwilling to share their techniques. The loose partnership requirements of contributing within Archimedes allow commercial companies to contribute their features or results sets without disclosing their techniques, enabling unfettered communication. The system also enables the rapid creation, storage, and download of specialized data sets for comparisons. One example of an interesting data set that Archimedes can create and store would be mammograms of high-density breasts for which MRI images are also available. Comparison of CBIR results is simplified by the storage of multiple results sets for images, and the ability to quantify the results sets.

There are several different types of research that go into developing an effective CBIR system for medical images. The primary research is done by radiologists, who perform the medical scans as well as provide diagnoses. Archimedes can help

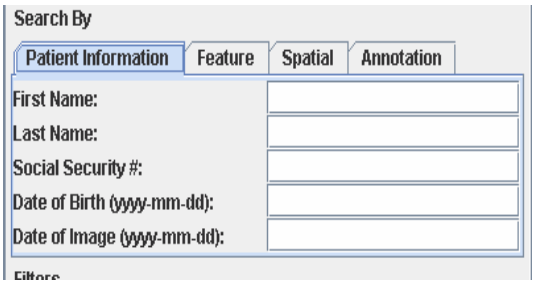


Figure 1. Archimedes Patient Information Search Panel. Searches over patient information can be done using the patient's name, date of birth, social security number, or the date an image was taken.

Figure 2. Archimedes Filter Panel. Results from searches can be filtered based on pathology, image and scanner type, weight range, and race.

radiologists by organizing their images, capturing their patient notes and digitizing their image annotations, speed up the analysis of experiments, and enable quick comparisons of different radiologist techniques such as comparing double-reading to single reading of mammograms. Web publication of research into radiologist techniques can be simplified using Archimedes' built in annonimizing and web publishing. The next stage of research is feature extraction and analysis. An example of this is the measurement of spiculation [13] as a feature used to aid in the detection of spiculated lesions. Archimedes can help with research into feature extraction by providing annotated images, comparison features, and comparison results sets, as well as data analysis and feature combination. Once the images have features and diagnoses associated with them, all of the pieces are available for research into CBIR techniques.

The rest of this paper is organized according to the steps in the process of creating a CBIR system. Section 2 discusses the use of Archimedes for data collection and analysis by radiologists. Section 3 reviews the unique feature upload, storage, and sharing capabilities of Archimedes. Section 4 describes how Archimedes is used to create and evaluate effective CBIR techniques, Section 5 discusses the design, and Section 6 draws conclusions and discusses future work.

2. Data Collection and Analysis

Archimedes is designed to be able to run and analyze double-blind studies of radiologist techniques. Archimedes can be configured so that multiple radiologists can annotate the same image under different conditions without viewing the biopsy-based ground truth or the other annotations. Their input can then be viewed by a user which can only view the doctor information after it has been automatically annonimized to maintain the integrity of the study. The data can be analyzed in several ways using the capabilities of Archimedes. A particular data set can be isolated using the dataset name in the image annotations. This data set can be filtered to show only particular types of cases, such as malignant cancers, normals, or benign cancers. A particular (annonimized) doctor's diagnosis can further refine

the results, showing the percentage of accurate diagnoses on different classes of images. The position of the actual cancer can be compared to the biopsied "truth" position using a spatial search, finding all of the cases where the doctor's diagnosis is within a specified distance of the "truth" position. The analysis can then be finished in a matter of hours. The results can then be stored in Archimedes, and the study can be published immediately by allowing Archimedes to enable guest user access.

Archimedes was originally designed to do double-blind research studies, but it also has an image analysis and patient records management tool that can be used by the medical research community. Radiologists can store and organize medical images such as x-rays, mammograms, CAT scans, MRIs, and any other image that is stored in a DICOM format. Radiologists can rapid retrieve images and patient records, and can also find patients with similar images, conditions, or annotations to compare treatment successes. The software archives the addition of markups and notations to images, as well as associating text and patient info with images.

For managing patient records and images, the primary tool is searching by patient information and text. For search by text in Archimedes, the medical professional is allowed to enter patient information (i.e. first name, last name, date of birth, etc.) into Archimedes. Once information is entered, they can use filters to further refine their search results. The searching options for patient information are shown in Figure 1, while the filtering options are shown in Figure 2. Searching by feature allows the doctors to specify feature parameters that they wish to see in the results and can be seen in Figure 3. This enables medical professionals to quickly find similar cases. The next type of search is an extension to searching by feature, specifying multiple features with defined spatial relationships between them as in Figure 4. The most useful of these types of search would specify a distance between features, for example to find areas that have the features of both spiculations and bright central cores indicative of spiculated lesions in mammograms. Archimedes also allows search over comments other doctors previously made about patients or images as shown in Figure 5.

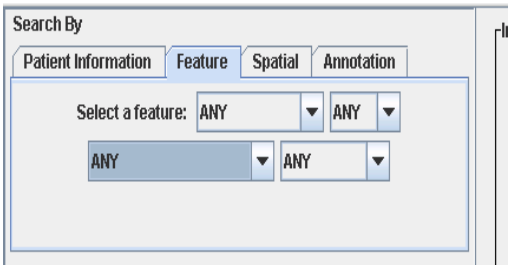


Figure 3. Searching by feature allows the doctors to specify feature parameters they wish to see in all the images in the return set.

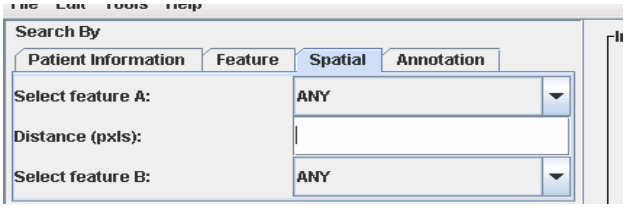


Figure 4. Doctors can use spatial searching when searching certain information about images. They can be looking for combinations of features in an image, or clusters of features in an image. The results will be displayed in the results list, allowing doctors to choose or further filter them.

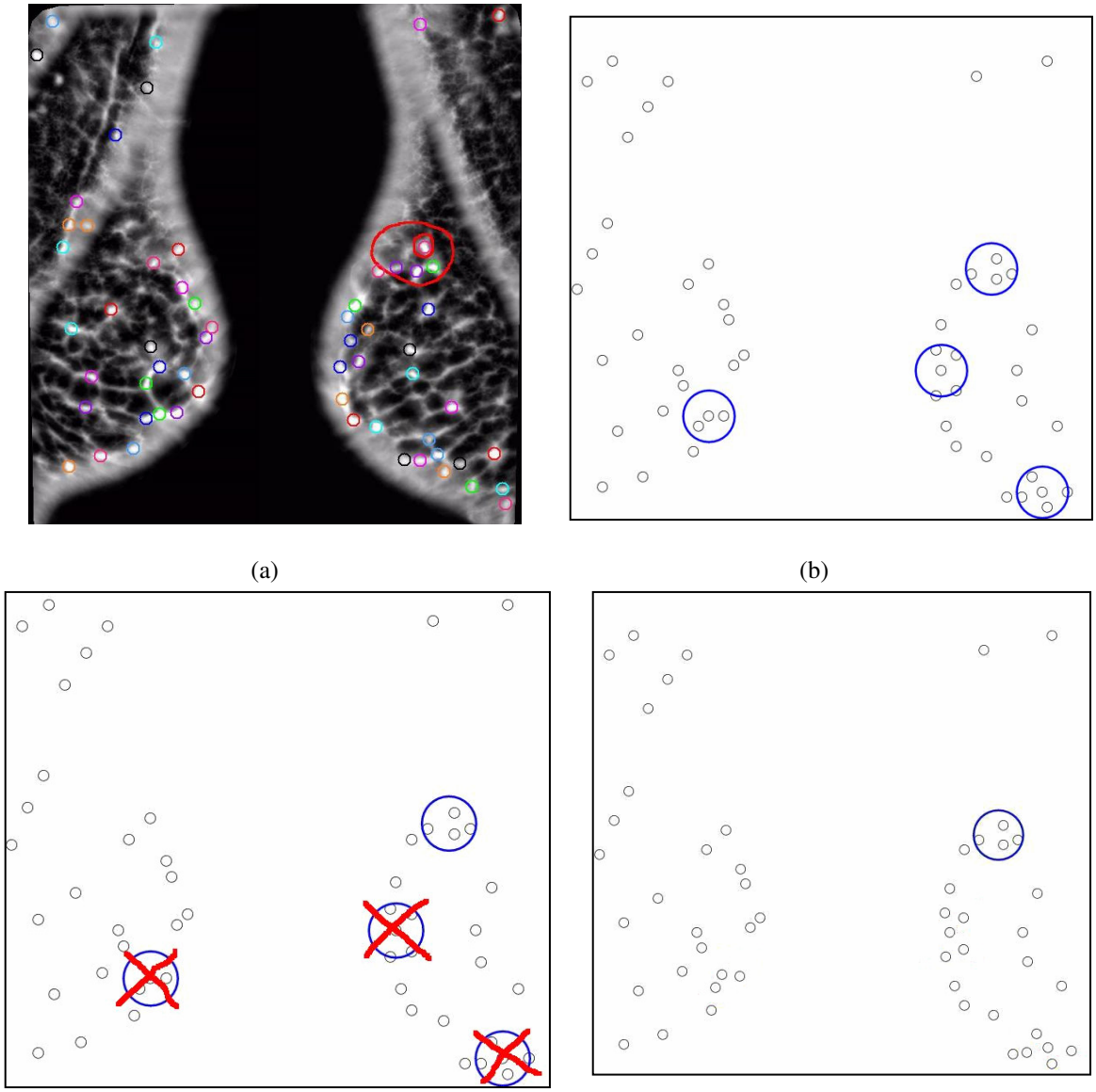
Figure 5. Annotation search allows doctors to search for comments other doctors previously made about certain images. Not only will this help in quickly determining information about a patient, but can also help doctors in understanding new patients' with similar problems or diseases. An example of a doctor using an annotation search would be for the words "biopsy recommended". This is similar to a search on yahoo over medical annotations.

This works like a primitive yahoo search over the text of the medical annotations.

Though the basic searching on patient information and annotations is included with minimal effort, the more advanced feature and spatial searching requires extra input. There are two options: capturing doctor input or getting permission to make images available for researchers to mark up the images. For example, getting the features like spiculation into the medical database may require making those images available to the researchers who specialize in measuring spiculation. Making the images available to researchers involves setting the permissions for a individual image or groups of images to public or semi-private, where images are available to the public at large or to a set group. The signed consent forms can also be stored in Archimedes as images. Capturing the doctor input required a viewing and input capture tool.

Medical professionals can view and manipulate images on the Archimedes system and tab through set of images. There is a zooming interface in order to focus in on interesting parts of the images. Point features can be inserted and described as overlays to the images. Text annotations can be entered. Multiple overlays can be captured for each image, allowing double reading of images. This type of detailed image information is essential for the design and evaluation of CBIR systems as well as computer-aided detection and diagnosis systems. The radiologists' diagnoses are captured and the data can be accessed remotely, thereby allowing tele-medicine applications to be run on the Archimedes platform.

The Archimedes system helps store, annotate, and retrieve data and images for the radiologists, who are the primary data collection agents in the development of a medical image CBIR system. Improving their work and giving them incentives to go through the paperwork necessary to share their images and diagnoses is one of the key steps in creating a collaborative environment to create CBIR.



(a) (b)

(c) (d)

Figure 6. An example mammogram image pair that might be stored in Archimedes is in (a). The small multi-colored circles are computer-generated features, while the larger red circle is the radiologist diagnosis. Archimedes spatial search can be used to find clusters of features, and this is shown in (b). However, there are still multiple noise clusters. Storing these clusters as features, we can use a spatial query again to find all of the clusters that are a certain distance away from the breast boundary, which eliminates three noise clusters as shown in (c) and leaves only the actual cancer as shown in (d). This is one of the ways that Archimedes can be used to combine features. The effectiveness of the feature combination can be evaluated using a spatial search again to find the features that are within a certain distance from the radiologist's biopsy-verified diagnosis.

3. Features

One of the main challenges in feature extraction is finding a large enough set of images of the same exact type of cancer in order to focus in on its particular characteristics. But there are a few large databases that do provide these images, for example with lung cancer images [4]. However, in order compare the effectiveness of one feature versus another on the same images, the comparison research has to be replicated. Archimedes eliminates this problem by allowing researchers to store their features in Archimedes for comparison along with the images from which they were extracted. The input, storage, and sharing of features is one of the design choices that make Archimedes unique. The open sharing of features makes comparisons possible without the need for inaccurate replication of older work, as well as enables research into feature combination both faster and more effective. Features can be combined using spatial search and then fused into a new feature type. An example of this is shown in Figure 6. For example, a spiculation feature can be combined with a bright central core feature for detecting spiculated lesions. The spatial search can also be used to quantify the effectiveness of the feature at predicting the position of the cancer by comparing the feature position with the biopsied “truth” cancer position.

Evaluating the effectiveness of features can be done using Archimedes extensive query capability. Finding cases when a “test” feature is near a “truth” feature can be done using the spatial query capability, as well as those “truth” features that are not near a “test” feature and vice versa. The spatial query is tunable with a variable input distance for greater flexibility. The image categories can be adjusted with the filters to isolate cases that are malignant, or other medically relevant characteristics.

Archimedes can also store and share features that are not associated with a particular position. The flexible design of the feature storage and upload make it capable of handling most features. A planned improvement is the handling of feature areas and feature volumes. Currently only point features are handled, but in order to handle segmentation data and other area features and upgrade is required and is expected in late 2007.

Features can be manually input through the Archimedes zooming interface, or loaded in bulk though an XML schema, with a small example:

```
<Patient>
    <Image>
        <Doctor>
            <Feature>
            </Feature>
            <Feature>
            </Feature>
            <Annotation>
            </Annotation>
```

```
        </Doctor>
    </Image>
</Patient>
```

Archimedes generates a skeleton schema for a user's selected group of images in order to facilitate upload and match the annonimized patient ID with the correct image. Features can contain pixel positions using <Xpos> and <Ypos>, but it is not required. Features can contain a number of values associated with them, and these values are uploaded with <ValueXName>, <ValueXType>, and <ValueX> for the Xth value. These values can be used in limiting feature searches.

Techniques that combine features could also be stored in Archimedes, and stored as features as well. Currently the spatial search can be used to combine features, but more complex approaches have to be done offline and uploaded as features. One planned upgrade is a learning package built into Archimedes that would simplify the development of classifications and analysis of medical images. This would allow radiologists to use Archimedes to explore relationships and use the learning package to optimize the approach.

The combination of advanced querying and feature sharing enables rapid analysis of features and combinations of features for CBIR and the comparison of computed features to "truth" features defined by a radiologist. By providing a platform for the analysis and comparison of features, Archimedes encourages the collaboration between researchers designing features as well as researchers building CBIR.

4. CBIR

The images, categorizations, and diagnoses provided by radiologists combined with features enable the exploration of CBIR in medical images through Archimedes. Typical CBIR approaches combine features into a feature vector and use a variety of techniques to determine the most accurate similarity measure. The categorizations of images, like evaluations of breast density or cancer type or malignancy, can be used to evaluate and to verify the effectiveness of CBIR at returning similar images.

As with features, it can be difficult to compare CBIR techniques without recreating the research of others. Archimedes allows the storage of result sets to simplify the comparison of different CBIR approaches. Currently the results sets are stored with the query image, but as images are added to the database over time the results set should change. The date needs to be added in order to prevent the comparison of newer images in competing CBIR approaches.

Currently Archimedes only stores CBIR results, but a planned upgrade would allow CBIR techniques to be stored and utilized within Archimedes as well. CBIR techniques that can be stored as a matrix operation on a feature vector can be stored and used as an index.

5. Design

The design of Archimedes had to take into account the sensitive nature of the data as well as the multitude of regulations coming to govern this field. The design focused on satisfying HIPPA regulations in the US while maintaining the ability to adapt to other regulations.

The program and database must maintain a high level of security due to privacy issues associated with maintaining patient sensitive medical information. The application is web-based for simplified deployment and tele-medicine uses, but this makes security more of an issue. Information transmitted to the front-end is encrypted via the AES encryption scheme. The images are also encrypted between the database and the front-end. All modifications during program use are monitored and logged by the system, and the viewing of the logs is limited to administrators.

Access to images can be either tightly controlled and private, public, or semi-private, while access to patient information is always tightly controlled and private. Administration is simplified through the use of groups, where semi-private images have groups of trusted researchers associated with them to help provide analysis. This is also helpful for administering research studies where the research group is set as the image default setting. An example of how groups simplify workflow is where the tech uploads the images into Archimedes, and then they are available to the radiologists in the appropriate group.

Archimedes is a three-tiered application including backend server, server logic unit, and web front end user interface. The server can run on any machine using a Unix, Linux, or Windows operating system that can support Java 1.5 and MySQL. Additional details on the design can be found in [15].

6. Conclusions and Future Work

We have created a secure web-enabled HIPPA-compliant database for the storage, retrieval, manipulation, and annotation of medical images and medical records for the development and evaluation of CBIR methods. The most unique quality is the ability to input, store, and share multiple feature sets and result sets for each image, thereby allowing greater flexibility for CBIR and allowing web collaboration in the development of CBIR. Each expert needed for the development of CBIR gains advantages in their individual work by collaborating in Archimedes, while improving the project overall. The advanced querying and feature storage capabilities provide rapid analysis and comparisons radiologist techniques, medical image features and CBIR techniques.

Future work on this project on the hospital applications includes changes to easily incorporate the OCR converted scanned forms for older patient data as well as associating a scanned permission form with images that are to be made public or semi-

private. A test site has been deployed and we are awaiting deployment at a larger hospital. On the other hand, future work on the research applications side includes support for area and volume features, spatial search on features with certain value ranges, and the incorporation of an arbitrary number of feature values.

Thanks

This project was developed in cooperation with the Software Engineering At Maryland (SEAM) cooperative at the University of Maryland. This cooperative matches project with outstanding senior level computer science and computer engineering students at the University of Maryland. Some of the students who helped with this project include Matt Fowle, Dan Ilkovich, Nima Negabhan, James Wren, Guilherme Bandeira, Duane Gilbert, Matt Weinstein, Hassan Shaukat, Sureshmi Wijewardena, Bernard Ng, Paul Carlson, Ratandeep Singh Achreja, Zvi Alexander Band, Jay Ming-Chie Liu, Pratik Mathur, Htin Kyaw Nyo, Kristofer Patrick Quinn, Obaid Siddiqui, Michael Andrew Tantino.

Bibliography

1. A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C.-R. Shyu, A. Marchiori, Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment, Radiology 228 (2003) 265-270.

2. A.A.T. Bui, R.K. Taira, J.D.N. Dionision, D.R. Aberle, S. El-Saden, H. Kangarloo, Evidence-based radiology, Academic Radiology 9 (6) (2002) 662-669.

3. J.P. Boissel, M. Cucherat, E. Amsallem, P. Nony, M. Fardeheb, W. Manzi, M.C. Haugh, Getting evidence to prescribers and patients or how to make EBM a reality, Studies in Health Technology and Informatics (95) (2003) 554-559.

4. L.P. Clarke, B.Y. Croft, E. Staab, H. Baker, D.C. Sullivan, National Cancer Institute initiative: Lung image database resource for imaging research, Acadademic Radiology 8(5) (2001) 447-50.

5. M. O. Guld, B. B.Wein, D. Keysers, C. Thies, M. Kohnen, H. Schubert, T. M. Lehmann, A distributed architecture for content-based image retrieval in medical applications, in: Proceedings of the International Conference on Enterprise Information Systems (ICEIS2001), Setubal, Portugal, 2001, pp. 299-314.

6. D. Keysers, J. Dahmen, H. Ney, B. B. Wein, T. M. Lehmann, A statistical framework for model-based image retrieval in medical applications, Journal of Electronic Imaging 12 (1) (2003) 59-68.

7. C. LeBozec, M.C. Jaulent, E. Zapletal, P. Degoulet, Unified modeling language and design of a case-based retrieval system in medical imaging, in: Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA), Nashville, TN, USA, 1998, 887-891.

8. T. M. Lehmann, M. O. Guld, C. Thies, B. Fischer, M. Keysers, D. Kohnen, H. Schubert, B.B. Wein, Content-based image retrieval in

medical applications for picture archiving and communication systems, in: Medical Imaging, Vol. 5033 of SPIE Proceedings, San Diego, California, USA, 2003, 109–117.

9. H.U. Lemke, PACS developments in Europe, Computerized Medical Imaging and Graphics 27 (2002) 111–120.

10. S.C. Orphanoudakis, C.E. Chronaki, S. Kostomanolakis, I2Cnet: A system for the indexing, storage, and retrieval of medical images by content, Medical Informatics 19 (2) (1994) 109–122.

11. S.C. Orphanoudakis, C.E. Chronaki, D. Vamvaka, I2Cnet: Content-based similarity search in geographically distributed repositories of medical images, Computerized Medical Imaging and Graphics 20 (4) (1996) 193–207.

12. B. Revet, DICOM Cook Book for Implementations in Modalities, Philips Medical Systems, Eindhoven, Netherlands, 1997.

13. M.P. Sampat and A.C. Bovik, Detection of spiculated lesions in mammograms, Proceedings of the 25th Annual International IEEE Conference on Engineering in Medicine and Biology Society, (1) (2003) 810–813.

14. C.R. Shyu, C.E. Brodley, A.C. Kak, A. Kosaka, A.M. Aisen, L.S. Broderick, ASSERT: A physician-in-the-loop content-based retrieval system for HRCT image databases, Computer Vision and Image Understanding (special issue on content-based access for image and video libraries) 75 (1/2) (1999) 111–132.

15. D. Tahmoush, H. Samet, A new database for medical images and information, in Proceedings of SPIE – Medical Imaging 2007: Image Processing, San Diego, CA, February 2007.

16. M.W. Vannier, E.V. Staab, L.C. Clarke, Medical image archives – present and future, in: Proceedings of the International Conference on Computer-Assited Radiology and Surgery, Paris, France, 2000, 565–570.