Towards Knowledge-Based Retrieval of Medical Images. The Role of Semantic Indexing, Image Content Representation and Knowledge-Based Retrieval.

Henry J. Lowe MD.[A]; Ilya Antipov BS [A]
William Hersh MD.[B] Catherine Arnott Smith MA, MILS [A]

[A] Center for Biomedical Informatics, University of Pittsburgh
[B]Division of Medical Informatics and Outcomes Research, Oregon Health Sciences University

ABSTRACT

Medicine is increasingly image-intensive. The central importance of imaging technologies such as computerized tomography and magnetic resonance imaging in clinical decision making, combined with the trend to store many "traditional" clinical images such as conventional radiographs, microscopic pathology and dermatology images in digital format present both challenges and an opportunities for the designers of clinical information systems. The emergence of Multimedia Electronic Medical Record Systems (MEMRS), architectures that integrate medical images with text-based clinical data, will further hasten this trend. The development of these systems, storing a large and diverse set of medical images, suggests that in the future MEMRS will become important digital libraries supporting patient care, research and education. The representation and retrieval of clinical images within these systems is problematic as conventional database architectures and information retrieval models have, until recently, focused largely on text-based data. Medical imaging data differs in many ways from text-based medical data but perhaps the most important difference is that the information contained within imaging data is fundamentally knowledge-based. New representational and retrieval models for clinical images will be required to address this issue. Within the Image Engine multimedia medical record system project at the University of Pittsburgh we are evolving an approach to representation and retrieval of medical images which combines semantic indexing using the UMLS Metathesuarus, image content-based representation and knowledge-based image analysis.

INTRODUCTION

The digital imaging revolution of the past 25 years has transformed medicine[1] . Technologies such as Computerized Tomography (CT) and Magnetic Resonance Imaging (MRI) have radically changed the way physicians diagnose and treat disease. The maturing of PACS[2] technology suggests that most, if not all, radiological data will be stored and retrieved digitally within the next decade. In addition, we are witnessing the emergence of digital imaging repositories in many disciplines such as pathology[3], Cardiology[4] and dermatology[5]. The more recent developments of functional imaging modalities such as Functional MRI and Positron Emission Tomography (PET) coupled with software to create three dimensional reconstructions based on two dimensional image data sets suggests that these technologies will also become an essential clinical decision support tool. Future clinical information systems will need to handle this heterogeneous imaging data sets if they are to effectively support patient care. Managing very large amounts imaging data[6], ensuring the network bandwidth required to deliver it to the point of care and developing database architectures that can seamlessly integrate both the textual and imaging components of the medical record are all major challenges that confront the designers of multimedia medical record systems. The current heterogeneity of medical imaging data formats is likely to become less of a problem as DICOM[7] is adopted as the universal medical image data format.

THE MEDICAL RECORD AS A DIGITAL LIBRARY

The patient record has long been an important database supporting clinical care, medical research and education. As the patient record increasingly becomes an electronic entity, encompassing digital imaging data and supporting secure access via the Internet, it can evolve into a virtual digital library supporting not only patient care but also innovative research and educational applications. This virtual database system will permit queries such as "Is the prevalence of pulmonary metastatic lesions in patients with Malignant Melanoma related to the anatomic site and/or stage of the primary lesion?" Currently such a query would require an expensive and time-consuming review of patient charts, radiology reports and surgical pathology documents. However the information required to answer questions such as this exists within a process linking external knowledge bases, human experts and the radiological and pathology imaging data itself. We know this to be the case because medical experts can

1091—8280/98/$5.00

analyze this imaging data and generate the textual reports needed to resolve the query described above. Imaging reports are therefore secondary data sources that attempt to link the important information components of primary imaging data and related external knowledge bases in much the same way that MeSH terms attempt to represent the important information concepts contained within a paper cited by MEDLINE. Secondary information representation is always problematic because it is dependent upon the expertise, focus, context and semantics of its creator. This is particularly the case with medical images where the human expert will often focus primarily on the data within an image that is relevant to the clinical question at hand. We believe that a new model for medical image representation is required which combines the information retrieval methods currently used to retrieve data from textual databases[8] with exciting new approaches to image content-based representation emerging outside of medicine[9]. We also believe that it will be essential to integrate knowledge-based representational schemas into this model if we are to effectively apply medical imaging data to clinical and research problems.

A MULTIDIMENSIONAL APPROACH TO MEDICAL IMAGE RETRIEVAL

The Image Engine project[10] is a medical informatics research effort at the University of Pittsburgh, funded by the National Library of Medicine's National Telemedicine Initiative, to prototype an integrated multimedia electronic medical record system. The system , under development since 1993, is currently in daily use at the University of Pittsburgh Cancer Institute where it provides a complete multimedia medical record for several hundred clinically active cancer patients. For each patient the system stores all radiology, pathology, clinical photography and EKG images. In addition each image is linked, in real-time using software agents, to imaging reports, clinical documents and laboratory test results stored in external repositories at the University of Pittsburgh Medical Center. At the time of writing , Image Engine contains approximately 25,000 clinical images related to almost 5000 clinical imaging procedures. We are therefore confronted with the problem of how one can support retrieval from a relatively large, heterogeneous medical image database. Fundamentally this is a problem of how one represents the information content of heterogeneous image sets and how one links this information to the traditional text-based data contained in the patient record and external knowledge bases. The approach proposed in this paper represents the preliminary results of our analysis of a best effort to begin addressing this issue.

We currently recognize four complementary approaches to describing the information related to medical images in the Image Engine database. These are:

1. Representation based on demographic and procedure information

2. Semantic representation based on analysis of related clinical documents

3. Direct image content-based representation

4. Knowledge-based representation

1. Representation based on demographic and procedure information

Many clinical images exhibit two universal attributes. They are usually images of a particular patient and they are created by a specific clinical procedure at a particular point in time. Therefore one important approach to image representation is based on using patient demographic data (name, date of birth, gender, patient ID etc.) and procedure data (procedure class, procedure name, CPT code, date/time/location of procedure and indications). This approach, currently supported by Image Engine, permits queries such as "Retrieve all CT scans of the chest performed after 1/1/1998 on patient John Doe, born 9/12/1943". Our preliminary analysis of image retrieval by users of Image Engine suggest that many images requested for patient care are successfully handled using this type of query. This retrieval approach exhibits excellent precision and recall performance when the database queries are limited to a particular patient or a set of patients that can be identified by demographics and or procedural information. The data required to support this type of query can be automatically extracted from the DICOM headers of many radiological images and/or the procedure ordering records used by most ADT and Radiology systems. At the time imaging data is entered into the database this extracted data is be used to automatically populate demographic and procedural indices that will facilitate retrieval of the relevant images. In addition, one can improve retrieval performance by using a standardized vocabulary to represent procedures. We currently use terms derived from NLM's Unified Medical Language System Metathesaurus to both represent procedures and to assist users in identifying procedures when creating database queries.

2. Semantic representation using imaging reports

Another attribute that most clinical images possess is that they are analyzed by a human domain expert, a radiologist for example, and this analysis results in a textual report that summarizes important information

contained within the image. Clinical image reports represent secondary information generated by a domain expert following analysis of the primary imaging data. In the Image Engine project we are currently attempting to represent the important concepts found in imaging reports using the NLM's Unified Medical Language System Metathesaurus. Our approach to this task is to use the SAPHIRE[11] indexing engine, developed by Hersh at the Division of Medical Informatics and Outcomes Research, Oregon Health Sciences University, as an automatic UMLS indexing system. The reports related to imaging data stored by Image Engine are parsed by SAPHIRE to produce a list of UMLS concept descriptors. Each concept so identified is represented in the image Engine database using the UMLS concept preferred name, semantic type and concept unique identifier. Our preliminary analysis of automated UMLS indexing of image reports, using the 1998 metathesaurus, suggests that for a wide range of radiology reports the UMLS can provide appropriate anatomical, procedural, disease and findings descriptors. Controlled vocabulary indexing has many advantages over "free text" retrieval by resolving the query ambiguities that often occur with uncontrolled text. To assist users in identifying the appropriate UMLS terms with which to create controlled vocabulary queries of the image database we have created, within Image Engine, a software agent-based system that accesses NLM's Knowledge Sources Server[12] over the Internet. Users access this system to map terms to controlled UMLS descriptors which they then apply to the construction of image queries.

Our preliminary analysis suggests that automated UMLS indexing of imaging reports can support retrieval of images independent of patient demographic data. This type of retrieval is most likely to be useful to researchers and educators who need to discover subsets of images that exhibit certain attributes. For example a query such as "find all abdominal CT Scans which show Renal Cell Carcinoma and liver metastases" or "find all microscopic pathology images of brain tumors in which cells exhibit Rosenthal fiber" can be effectively handled using this approach. The use of UMLS data structures such as semantic type, related terms and hierarchical relationships suggests that this retrieval model can support powerful retrieval tools such as term explosion and semantic filters. Precision and recall performance will be dependent upon a number of factors: (i) the ability of the indexing engine to identify concepts at a sufficient level of granularity; (ii) the indexing engine's capacity to filter the negative references frequently encountered in clinical reports (e.g. "there is no evidence of liver metastases"); (iii) whether the scope and granularity of the UMLS is sufficient to represent concepts identified by the indexing engine and (iv) whether the retrieval system can assist the user in identifying the appropriate UMLS descriptors for inclusion in image retrieval queries.

Preliminary results suggest that SAPHIRE can identify greater that 75% of the important concepts in image radiology reports and map them to appropriate UMLS descriptors, though probably this approach will emphasize retrieval recall over precision. As stated earlier, based on hand-indexing of imaging reports the 1998 UMLS Metathesaurus appears to contain terms of sufficient granularity to cover many important concepts found in radiology and pathology imaging reports.

3. Image Representation based on image feature extraction and segmentation

To answer a query such as "Find all chest x-rays which show enlargement of the left ventricle" one could use the semantic approach outlined above. The success of this representational approach assumes that the person generating the report actually recorded the features of interest in sufficient detail and used language that the indexing engine can recognize. However, in our experience, this is not always the case. For example, if a chest x-ray is ordered by a physician to determine if a patient has pneumonia and if that patient is known to have a history of cardiac enlargement then the radiology report may only refer to this finding in passing (e.g. "cardiomegaly is again noted"). The indexing engine would detect the term cardiomegaly but without additional knowledge that defines left ventricular enlargement as a type of cardiomegaly, it is unlikely that this image would be retrieved using a semantic search. In our experience this phenomenon is not uncommon, as image reports tends to be contextual and will frequently focus on the problem context that generated the imaging request in the first place. Image retrieval queries of this type will require a combination of semantic indexing (to define the subset of images exhibiting heart enlargement), image content-based indexing (which attempts to identify images (a) exhibiting heart enlargement not identified by semantic indexing and (b) pure left ventricular enlargement which may not be recognized as related to Cardiomegaly by a purely semantic representation).

To represent information within images and with sufficient granularity to ensure adequate recall performance one must develop strategies for automated identification of features within the image itself. Considerable research on automated and assisted feature identification has been carried out in both non-medical domains and within biomedicine. For example, the Query by Image Content (QBIC)[13] system analyses images and represents the image in a multidimensional information space using image

vectors related to attributes such as colors, textures and shapes. QBIC uses this "content indexing" to support similarity searches in which users iteratively request image sets that are similar to a seed image presented by the system. QBIC also allows a user to create rough outline sketches using shapes and colors to describe properties of the images to be retrieved. Similar work on retrieving radiological images using a sketch-based interface has been described.[14]

A more promising, complementary approach to image content representation involves the use of segmentation algorithms to identify important features within an image. Segmentation is the process of determining a region or feature of interest in a scene. Thresholding labels image pixels based on intensity value in relation to a threshold value. In its simplest form applying a single threshold value will result in a binary image where "object" segments are differentiated from the background. The "flood-fill" method, which starts with a single object pixel and repeatedly add adjacent pixels whose values are within some given threshold of the original pixel can also be used to segment imaging data into feature objects. This type of technology is increasingly being used with biomedical images[15,16] to identify anatomic features and to attempt to identify whether there is significant deviation from a features normal location, shape or size.

4. Knowledge-based approaches to medical image retrieval

The interpretation of medical images is inherently knowledge-based[17]. For example, when a physician looks at a chest x-ray which demonstrates an area of increased density in the lower part of the left lung she may execute the following steps:

A. Identify the modality class - "Its a radiograph"

B. Identify the image type - "Its a chest x-ray"

C. Identify the anatomic field of view - "Its an anteroposterior view of the thorax"

D. Identify the major anatomic segments - "Heart, Left Lung etc."

E. Compare each segment to an internal norm - "Increased density in the left lower lobe"

F. Search knowledge base for possible causes of this finding

G. Use other imaging or history data to filter output of step F.

This analysis tree clearly demonstrates how the successful completion of each step is dependent upon the successful completion of the prior step. While "gestalt image interpretation" is more likely to be the case with domain experts, representing the process as a network allows one to break down the interpretive process into smaller steps which may be computable. Knowledge-based image analysis can aid significantly in this process when used for image representation rather than computer-assisted diagnosis. Recapitulating these steps demonstrates how a knowledge-based image representation system might operate.

A. Imaging modality, type and field of view identified from DICOM header: *Modality is radiograph, procedure in chest radiograph. View is anteroposterior*

B. Knowledge base indicates anatomic region is Thorax: *If the procedure is chest x-ray then the anatomic region is thorax. If the view is anteroposterior chest x-ray then AP view*

C. Knowledge base represents the main anatomic features in AP view of Thorax: *Heart, Left Lung, Right Lung, Mediastinum, Trachea etc.*

D. System uses segmentation to identify each important feature: *Knowledge base can initialize segmentation by "pointing" the segmentation algorithm into a region of the image where feature normally occurs.*

E. Each identified feature is compared to a reference for density, size, shape and location.: *Features can be further segmented into "parts" e.g. upper, middle and lower lobes of the right lung.*

F. Deviation from these reference values represented as vectors in database.: *Lower lobe of right lung contains a large area of increased density (higher than average pixel values) approximating to the usual area of the lower lobe itself.*

G. Vector representation linked to entry in disease database: *Increased density in the right lower lobe only could indicate pneumonia.*

CONCLUSION

Each of the four approaches that we have described for medical image representation and retrieval is complementary. In some situations a single strategy may be sufficient to retrieve the required image (e.g. demographic and procedural searches). However if we are to realize the research and educational potential of large heterogeneous image databases it is very likely that a combination of these approaches will be required. In our work we have already recognized the limitations of text-based image retrieval and we are now moving to explore how

knowledge-based image content representation can complement textual indexing.

ACKNOWLEDGMENTS

Dr Lowe's, Mr. Antipov's and Dr. Hersh's work on the Image Engine project is supported by National Library of Medicine National Telemedicine Initiative Contract number N01-LM-4-3507. Ms. Arnott Smith is supported by National Library of Medicine Medical Informatics Training Grant number 5-T15-LM07059.

http://www.cml.upmc.edu

References

[1] Lomas DJ, Dixon AK
Radiology. 100 years and running: the medical imaging revolution.
Lancet 1995 Dec 23;346 Suppl:S20

[2] Honeyman JC, Frost MM, Huda W, Loeffler WPicture archiving and communications systems (PACS).
Curr Probl Diagn Radiol 1994 Jul;23(4):101-158

[3] O'Brien MJ, Sotnikov AV
Digital imaging in anatomic pathology.
Am J Clin Pathol 1996 Oct;106(4 Suppl 1):S25-S32

[4] Thomas JD, Nissen SE
Digital storage and transmission of cardiovascular images: what are the costs, benefits and timetable for conversion?
Heart 1996 Jul;76(1):13-17

[5] Kenet RD
Digital imaging in dermatology.
Clin Dermatol 1995 Jul;13(4):381-392

[6]Honeyman, Huda, Frost.
Picture Archiving and Communications System Bandwidth and Storage Requirements
Journal of Digital Imaging Vol 9, No 2 (May) 1996

[7] Bidgood WD Jr, Horii SC, Prior FW, Van Syckle DE
Understanding and using DICOM, the data interchange standard for biomedical imaging.
Am Med Inform Assoc 1997 May;4(3):199-212

[8]Hersh WR
Information Retrieval: A Health Care Perspective
Springer-Verlag, 1996

[9]Dimitrova N, Golshani F
Video and Image Content Representation and Retrieval. The Handbook of Multimedia Information Management. Editors: Grosky W, Jain R, Mehrothra R. Prentice Hall 1997. Pages 95-138

[10]Lowe HJ, Walker WK, Polonkey SE, Jiang F, Vries JK, McCray A
The Image Engine HPCC Project. A Medical Digital Library System using Agent-Based Technology to Create an Integrated View of the Electronic Medical Record.
Proceedings of the Third International Conference on Research and Technology Advances in Digital Libraries (ADL 96). Washington DC, May 1996; IEEE; Pages 45-56.

[11]Hersh WR, Leone TJ,
The SAPHIRE server, Proceedings of the 19th Annual Symposium on Computer Applications in Medical Care, 1995, 858-862.

[12]McCray AT, Razi AM, Bangalore AK, Browne AC, Stavri PC
The UMLS Knowledge Source Server-a Versatile Internet-Based Research Tool
Proc AMIA Fall Symp 1996;164-8

[13] Query by Image and Video Content: The QBIC System
Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q et al
IEEE Computer, September 1995, 23-31

[14] Robinson GP, Tagare HD Duncan JS Jaffe CC
Medical image collection indexing: shape-based retrieval using KD-trees
Comput. Med. Imaging Graph; vol.20, no.4; July-Aug. 1996; pp. 209-17

[15]Betal D, Roberts N, Whitehouse GH
Segmentation and numerical analysis of microcalcifications on mammograms using mathematical morphology.
Br J Radiol 1997 Sep;70(837):903-917

[16]Saiviroonporn P, Robatino A, Zahajszky J, Kikinis R, Jolesz FA
Real-time interactive three-dimensional segmentation.
Acad Radiol 1998 Jan;5(1):49-56

[17]Hemant D. Tegare; C. Jaffe C; Duncan J
Medical Image Databases: A Content-Based Retrieval Approach
JAMIA, Vol. 4, No. 3, Pages 184-198