Journal of Digital Imaging

A Visual Query-by-Example Image Database for Chest CT Images: Potential Role as a Decision and Educational Support Tool for Radiologists

Giuseppe Sasso, MD, PhD,^{1,2,3,6} Hugo Raul Marsiglia, MD,⁴ Francesca Pigatto, MD,^{2,4} Antonio Basilicata, MD,^{1,2} Mario Gargiulo, MD,^{1,2} Andrea Francesco Abate, BSc, PhD,⁵ Michele Nappi, BSc, PhD,⁵ Jenny Pulley,^{2,3} and Francesco Silvano Sasso, MD¹

Primary reading or further evaluation of diagnostic imaging examination often needs a comparison between the actual findings and the relevant prior images of the same patient or similar radiological data found in other patients. This support is of clinical importance and may have significant effects on physicians' examination reading efficiency, service quality, and work satisfaction. We developed a visual query-by-example image database for storing and retrieving chest CT images by means of a visual browser Image Management Environment (IME) and tested its retrieval efficiency. The visual browser IME included four fundamental features (segmentation, indexing, quick load and recall, user-friendly interface) in an integrated graphical environment for a user-friendly image database management. The system was tested on a database of 2000 chest CT images, randomly chosen from the digital archives of our institutions. A sample of eight heterogeneous images were used as queries and, for each of them a team of three expert radiologists selected the most similar images from the database (a set of 15 images containing similar abnormalities in the same position of the query). The sensitivity and the positive predictive factor, both averaged over the 8 test queries and 15 answers, were respectively 0.975 and 0.91. The IME system is currently under evaluation at our institutions as an experimental application. We consider it a useful work-in-progress tool for clinical practice facilitating searches for a variety of radiological tasks.

KEY WORDS: Database, lung, computed tomography

THE PAST 20 YEARS have seen tremendous changes in medical imaging techniques: new modalities and protocols are rapidly expanding the available digital image data. Advances in imaging techniques have been mirrored by advances in the use of

computers to extract and interpret image data: helping clinicians to make effective use of this superabundance of information is a key aim of medical software research.

Primary reading or further evaluation of diagnostic imaging examination often needs a comparison between the actual findings and the relevant prior images of the same patient or similar radiological data found in other patients. This important clinical support can have significant effects on physicians' examination reading efficiency, service quality, and work satisfaction.

Medical documents are generally represented by text, whereas radiological documents are space-occupying images: this means considerable differences in archive management and related solutions. With conventional x-rays films and paper documents, making diagnostic images and

¹From the Diagnostic Imaging and Radiotherapy Department, Second University of Naples, Italy.

²From the S.G.C. Diagnostic Imaging and Radiotherapy Centre, Naples, Italy.

³From the Dorset Cancer Centre, Poole Hospital NHS Trust, Poole, Dorset, UK.

⁴From the European Institute of Oncology, Milan, Italy.

⁵From the Mathematics and Computing Department, University of Salerno, Baronissi, Salerno, Italy.

⁶From the Institute of Oncology, Townsville Teaching Hospital, Queensland, Australia.

Correspondence to: Giuseppe Sasso, MD, PhD, The Townsville Hospital, PO Box 670, Townsville, 4810, Queensland, Australia, tel: +61 (0) 7 4796 1626; fax: +61 (0) 7 4796 1621; e-mail: giuseppe_sasso@health.qld.gov.au

Copyright © 2004 SCAR (Society for Computer Applications in Radiology)

Online publication 28 October 2004

doi: 10.1007/s10278-004-1025-3

Table 1. IME System's Fundamental Features

Image features extraction by image processing	Edge detection and segmentation (hot spots and canonical objects extraction by means of an entropy-based method)
Image content representation and indexing	k-d trees-based data structure for index management (the spatial relationships among objects are the primary keys to index the images)
Storage and retrieval methodologies	Efficient spatial access to images by exact and/or inexact match (query by pictorial)
User-friendly interface	Graphical environments and visual query languages

study reports available for clinical purposes is time-consuming and space occupying for archiving. These additional costs and the risk of document unavailability or degradation are unacceptable, especially in situations emergency.

New diagnostic techniques such as ultrasound (US) and computed radiography have computed tomography (CT), magnetic resonance imaging (MRI) successfully replaced conventional film in data acquisition, image display, and interpretation. Because images are digitally acquired, and thanks to computer advances, we can now manage the whole of diagnostic imaging data with a Picture Archiving and Communication System (PACS).

The PACS coupled to patient data stored in a Radiology Information System (RIS), greatly reduces the perceived problem of image non-availability. Moreover, inexpensive and convenient databases are available for storage of digital radiological images in multiple formats or to create teaching files.^{1,2}

In general, PACS and image database workstations must be user-friendly, and physicians need to upgrade their knowledge of new techniques, modifying their working procedures.

The usefulness and understandability of a database application is considerably enhanced when the information is represented in multimedia forms, such as images, audio, and video annotations.³ The application of a large image database system can be considerably improved if non-expert users are provided with an intuitive and simple visual interface allowing easy image manipulation.

The purpose of this study is to implement a visual query-by-example image database for storing and retrieving chest CT images by means of a visual browser Image Management Environment (IME) and to test its retrieval efficiency as a means of assessing the system's performance

MATERIALS AND METHODS

System Characteristics

The IME system is an experimental environment for content-based access image management. It supports any application domain where the primary keys indexing the images are the spatial relationships among the objects. It incorporates a user-friendly visual interface, conceived as a typical Windows application, that merges the query-by-pictorial-example and query-by-sketch paradigms. A palette includes a set of anatomic objects for each human body region; these objects (i.e., right lung, left lung, spine, aorta, nodule, etc.) are taken from a standard CT scan and icon-represented by means of their contour.

A palette of the most common abnormality can be also included and upgraded according to the user's preference, adding multiple different visual objects.

The IME system includes four fundamental features in an integrated graphical environment: (1) image features extraction by image processing (edge detection and segmentation); (2) image content representation and indexing; (3) storage and retrieval methodologies for efficient spatial access to images by exact and/or inexact match; (4) a user-friendly interface (graphical environments, visual query languages, etc.) (Table 1).

The process of automatic storage consists of sequential contour extraction, image indexing, and fast data storage. The segmentation includes automatic contour extraction and evidence of abnormal parameters. The interface is Windows-based with a simple visual query toolbar to recall the main characteristics of the image (in this case CT lung images). The system analyzes the image content and assesses the object's spatial relationships as primary key.

The image-processing tools integrated in the IME system allow hot spots and canonical objects to be extracted by means of an entropy-based method for segmentation and edge detection.⁴ Before the image is stored in the database, it must be processed to extract the key features to be used for indexing. This is accomplished by assigning the correct meaning to the patterns in the image and selecting the significant ones. The system applies an image-indexing methodology based on *virtual images*⁵ that provide an efficient and reliable content-based access and similarity retrieval.

To achieve a faster search, IME is based on k-d-tree spatial access structures, offering demonstrated efficiency and gain with sequential scanning for comparing a large collection of image description data.⁶ The Image DataBase (IDB) was built from 2000 randomly chosen study sets with a single representative image selected for each study and

Fig 1. A lung CT scan with a suspect nodule: edge detection and query-by-example.

eight typical queries randomly repeated. The percent of abnormal images included in the database was 30 (600 pathologic images). The hardware has been constantly updated since 1997 and computing time, improving as the IDB grows, was confirmed from the first runs of the IME prototype system, on a Pentium 166 MHz processor PC in a standard configuration in 1991. The system now runs on a state-of-the-art PC.

This article describes the experimental application of IME to a "real-world domain" of CT images of the chest, a typical homogeneous image database system, because all the diagnostic images of a given human body district have approximately the same contents. As a case study, we have implemented a diagnostic database of chest CT images and evaluated its performance in terms of efficiency, mean response time for retrieval, and effectiveness, number of false alarms, and number of false dismissals.

The rationale for IME is to help the physician formulate a diagnosis according to the basic rules of the diagnostic imaging process: (1) identify abnormalities; (2) consider their spatial relationships to reference organs; (3) analyze such morphological and geometrical features as opacity, shape, symmetry, roundness, and area, among others; (4) evaluate the image's semantic contents and retrieve related images, by similarity, from a mnemonic database or a very large digital or analogical archive; (5) formulate a diagnosis based on personal knowledge and previously examined patients.

Figure 1 shows a typical IME window where the CT scan, namely the ANC10. BMP sub-window, is loaded in the upper-left corner. The results of its segmentation are shown in the ANC10. SEG sub-window. Besides these descriptors, a virtual representation of the image is automatically computed to describe the spatial contents of the real image and to allow content-based similarity retrieval. This

process does not require human intervention and currently takes less than 10 seconds per case on the latest workstation. All this collected information is in fact the *index* of the image. Figure 1 also shows an IME sub-window, namely the Query window, where a user is composing a query starting from a segmented image and using the iconic palette of the environment.

In this palette the iconic features of the organs commonly displayed in the chest CT scan, such as spine, left and right lung, etc., are represented. As previously stated, records of alphanumeric descriptors are automatically extracted from the image, such as spatial relationships, morphological and geometrical measures of each hot spot. Taking into account the spatial relationship among the structures displayed in the query window, all the images in the database with a similar abnormality in the approximately same position are retrieved.

Figure 2 shows the IME window that displays the results of the previous query, which include a set of retrieved images and their associated similarity degrees with reference to the query. Each image has a similar abnormality approximately in the same location of the image of Figure 1, with a similarity degree greater than 90% for all images. It is worth noting that no false alarms were retrieved—i.e., images not showing any abnormalities or images with abnormalities in different positions (e.g., within the left lung). Besides comparing their contents with those of the original image to formulate a diagnosis, the retrieved images can be used to formulate new queries.

Statistical Analysis

The *recall* and *precision* rates were used as evaluation criteria to measure the retrieval system's efficiency and as-

Fig2. The set of images resulting from the query in Figure 1.

sess the performance of IME: the recall is the system's capability to retrieve all relevant images (sensitivity); while the *precision* is the system's capability to retrieve only relevant images (positive predictive value). To relate the operating characteristics of medical tests into precision and recall (the terms more heavily used in document retrieval research, web mining, etc.), an equivalence among them and the parameters more used in Bioinformatics, Computer Medicine, evidence-based medicine, and other life sciences was established. This equivalence was assessed on the basis that both groups of terms measure comparable performance aspects in their respective areas (*Precision = Positive Predictive Value*; *Recall = Sensitivity*).^{7,8}

The positive predictive value (PPV) of the system (otherwise known as precision) was tested according to the following formula: $PPV = TP / (TP + FP)$; where the following parameters were defined: True Negative (TP) = images correctly identified as correlated to the query (same kind of disease). False Positive (FP) = images incorrectly identified as correlated to the query.

The sensitivity (S) of the system (otherwise known as *recall rate*) was tested according to the following formula: $S = TP / (TP + FN)$, where, in addition to the previously described TP, the following parameter was defined: False Positive (FN) = images incorrectly identified as non-correlated to the query (correlated images not recalled by the system).

The database of our case study was compounded by 2000 CT study covering a wide selection of pathological findings in different body areas (head, neck, chest, abdomen, and pelvis).

Eight sample images, (query), characteristic of different kinds of diseases, were selected by a team of three senior consultant radiologists. For each of these images, a set of 15 images were identified as better correlating answers (true

positive, TP) according to the shape (area, perimeter, symmetry, roundness, etc.) of the abnormality found and the spatial relationship to the standard objects previously described (lung, aorta, vertebral body, etc.).

RESULTS

We tested the IME system on a database of 2000 chest CT studies randomly selected from the radiological archives of our institutions and classified according to their different pathologies. The images, relative to the same human body area (chest), contain approximately the same objects in a comparable position; thus the main relevant difference between the images was related to the presence or absence of a similar abnormality, caused by several possible lung pathologies.

The system was tested with the previously described 8 queries. For each, an answer set of 15 images was required with the following results: The sensitivity (recall rate), averaged over 8 test queries and 15 answers for each query, was 0.975. Moreover: The median PPV over the entire 8 queries was 0.91 (91%, a mean of 1.35 FP for every 15 answer images). The highest PPV was 1 (100%, no FP) resulting from 3 queries. The lowest PPV was 0.8 (80%, 3 FP) resulting from 1 query.

If we consider the position (1st, 2nd, 3rd and 15th) in which the FP was re-called, which obviously was not taken into account by the PPV, and which represents the degree of similarity between the query and the answer images (the image re-called in the 1st position has a better degree of similarity than the one re-called in the 10th position), the results were as follow: Mean FP position = 9th mean value over 15 images. Highest FP position (worst re-call) = 3rd, in only one query. Lowest FP position (better re-call) = 14th, in 4 queries.

A last test was conducted to evaluate in the 5th query, in which the sensitivity was less than 1 (100%). The dimension of the answer set was extended to more the 15 images (16 or more) in order to assess the lowest number of re-called images needed to achieve the 100% sensitivity. The results were as follow: Mean number of re-called images needed to achieve the 100% sensitivity = 17 (mean over 5 query). Best score = 16 (2 query). Worst score = 19 (1 query).

DISCUSSION

The use of computers has yielded new diagnostic techniques (US, CT, MRI, and computed radiography) that can successfully replace conventional film in data acquisition, image display, and image interpretation.

Thanks to PACS, we can now manage the whole of the diagnostic database because images are acquired, coupled to patient data stored in Radiology Information System (RIS), sent to display workstations, and finally sent to the archives. Thus, PACS has greatly reduced the perceived problem of image non-availability, and automated image display increases efficiency in image interpretation and remedies the normally cluttered presentation environment improving physician productivity.⁹

An image database is important for research on digital imaging, such as image processing, image compression, image display, PACS, and computer-aided diagnosis. To date a number of researchers have developed digital image databases, and efficient prototypes are already available for mammography and chest radiography.¹⁰

Although the use of PACS has improved the workload management in radiology depart-

ments, the system still needs to be customized for the physician to be used optimally. Workstations must be user-friendly, with simultaneous display of more images. At the same times, physicians need to expand their knowledge of new techniques to modifying obsolete working procedure and maximize new ones.

Interface design is a key to the efficient use of medical workstation because it is the primary site of interaction between the physician and environment; it should be designed to make the system easy to use. When the interface is poorly designed, it is cumbersome and slow, causing frustration, requiring additional user time, and increasing labor costs of ancillary personnel. Though researchers have evaluated interface design¹¹⁻¹⁴, even when it is clear that the design of a diagnostic workstation can affect physician routine work efficiency, there is a substantial lack of data concerning how a system's characteristics affect radiographic interpretation, workflow, and physician performance. The automated image display method improves efficiency and simplifies the entire interpretative process.

Although the present study did not directly evaluate diagnostic accuracy, a greater focus on the cognitive process of data comparison may enhance diagnostic accuracy as well. Additional savings in interpretation time may be gained from use of the automated display instead of the work list display or manual film hanging; of particular value is the ability to display uninterpreted images first and then provide comparison studies sequentially in reverse temporal order.

In database applications the usefulness and understandability of a system is considerably enhanced if the information is represented in multimedia formats, such as images, audio, and video annotations.³ In particular, for applications where multimedia objects naturally play a central role, the usability of a large image database system can be considerably improved if non-expert users are provided with an intuitive and simple visual interface allowing for an easier image manipulation. Moreover, several applications of image databases require content-based access, but this kind of "raster image" is too complex and too large to be described by traditional indexing methods,

based on user-supplied alphanumeric descriptors. Unlike alphanumeric text, because of their size and internal complexity, medical images usually require a long time to be accessed, processed, stored, retrieved, and distributed. Thus, an electronic database should meet the requirement of effectively handling the above time-consuming operations.¹⁵

The PACS and image database management have become a basic research topic in computer science as new techniques are developed to provide quick delivery and easy access to images, video, text, and associated information, using appropriate representations for the intended tasks.^{16,17} To index an image, we can associate it with two kinds of descriptors: explicit information about its contents (in textual form) and implicit information related to the spatial arrangement of its pictorial elements and their morphological and geometrical measures. Then icons can be used as picture indexes, maintained according to an iconic indexing methodology.¹⁸⁻²⁰ With such a Query-by-Pictorial-Example approach,^{21,22} the index of an image is an iconic image itself. It represents the visual information contained in the image in a suitable form for different levels of abstraction and management. An efficient image database should integrate an image management technique (to access information uniformly and consistently according to a powerful database model) and a spatial image indexing methodology that allows efficient content-based access. It should also offer a user-friendly visual interface to help users store and retrieve images by means of a visual browser that easily navigates the database. The Image Management Environment (IME) consists of an integrated visual environment for loading digitized images, processing them to extract suitable spatial indexes, and storing and retrieving them easily thanks to a user-friendly visual interface.

It is worth noting that most of the existing (prototype and commercial) systems²³⁻²⁶ address only a subset of the previous requirements. A user-friendly visual interface lets users issue store-retrieve commands by simply editing real and iconic images by means of a palette of image processing tools and a palette of graphical objects.

We have presented our experience with an IME system,²⁷ a prototype environment for managing image databases applied on a medical database including a large number of lung CT images. The IME is composed primarily of a user-friendly management environment for handling image storage and retrieval by content. We have also illustrated the basic techniques developed to implement the image-processing tools, the iconic indexing of images in the system, and the SAM based on k-d-trees. The user-friendliness of the system is guaranteed by a visual browser following the query-by-pictorial-example paradigm. We ran several experiments to prove that our technique is both efficient and reliable in terms of false alarms and false dismissals. We consider it a potentially accurate and useful tool for clinical radiological practice, helping develop image databases that are easily searched for education, clinical assistance for radiologists, and research.

Acknowledgments

The authors are grateful to Dr. P. Vagnetti and Dr. A. Mendoza for their very helpful comments and suggestions in preparing this paper; to Dr. F. De Nisco for his kind and useful collaboration in this study; and to Prof. G. Tortora and Prof. M. Tucci and their colleagues working in research at the Mathematics and Computing Department, University of Salerno, Italy; where the IME was conceived and realized.

REFERENCES

1. Tran, THD, Roach, NA, O'Kane, PL, Thune, M: Creating a digital radiographic teaching file and database using a PC and common software. *AJR Am J Roentgenol* 175:325-327, 2000
2. Maldjian, JA, Listerud, J: Automated teaching file and slide database for digital images. *AJR Am J Roentgenol* 175:1249-1251, 2000
3. Sinha, U, Dai, B, Johnson, DB, et al: Interactive software for generation and visualization of structured findings in radiology reports. *AJR Am J Roentgenol* 175:609-612, 2000
4. Di Ruberto, C, Nappi, M, Vitulano, S: Different methods to segment biomedical images. *Pattern Recogn Lett* 18(11):1125-1131, 1998
5. Petraglia, G, Sebilio, M, Tucci, M, Tortora, G: A normalized index for image databases In: Chang, SK, Jungert, E, Tortora, G *Intelligent Image Database Systems*. World Scientific Publishing, Singapore, 1996, pp 43-69

6. Bentley, L, Friedman, JH: Data structures for range searching. *ACM Comput Surv* 11(4):397-409, 1979
7. <http://www.cs.helsinki.fi/u/borras/performance.html>
8. Salton, G, McGill, MJ: *Introduction to Modern Information Retrieval* New York: McGraw-Hill, 1983
9. Gale, DR, Gale, ME, Schwartz, RK, et al: An automated PACS workstation interface. A timesaving enhancement. *AJR Am J Roentgenol* 174:33-36, 2000
10. Shiraishi, J, Katsuragawa, S, Ikezoe, J, et al: Development of a digital image database for chest radiographs with and without a lung nodule. *AJR Am J Roentgenol* 174:71-74, 2000
11. Dreyer, KJ, Meththa, A, Johnson, KM, et al: The primary interpretation workstation: information beyond image data. *J Digit Imaging* 11(s2):16-29, 1998
12. Strickland, NH, Allison, DJ: Default display arrangements of images on PACS monitors. *Br J Radiol* 68:252-260, 1995
13. Kolodny, GM, Raptopelous, V, Simon, M, et al: A low-cost, full-function picture archiving and communication system using standard PC hardware and the traditional 4-over-4 display format. *AJR Am J Roentgenol* 172:591-594, 1999
14. Beard, D: Designing a radiology workstation: a focus on navigation during the interpretation task. *J Digit Imaging* 3:152-163, 1990
15. Jungert, E, Chang, SK: An image algebra for pictorial data manipulation. *Comput Vis Image Und* 58(2):147-160, 1993
16. Narasimhalu, AD: Multimedia databases. *Multimedia Syst* 4:226-249, 1996
17. Schnep, A, Du, DHC, Ritenour, ER: Building future medical education environment over ATM networks. *Commun ACM* 38(2):54-69, 1995
18. Chang, SK, Shi, QY, Yan, CW: Iconic indexing by 2-D string. *IEEE T Pattern Anal* 9(3):413-428, 1987
19. Chang, SK, Yan, CW, Dimitroff, DC, Arndt, T: An intelligent image database system. *IEEE T Software Eng* 14:681-688, 1988
20. Lee, SY, Hsu, FJ: 2D C-string: a new spatial knowledge representation for image database systems. *Pattern Recogn* 23:1077-1087, 1990
21. Chang, SK, Hou, TY, Hsu, A: Smart image design for large image database. *J Visual Lang Comput* 3:323-342, 1992
22. Chang, SK, Hsu, A: Image information systems: where do we go from here?. *IEEE T Knowl Data En* 4(5):431-442, 1992
23. Marsico, M De, Cinque, L, Levialdi, S: Indexing pictorial documents by their content: a survey of current techniques. *Image Vision Comput* 15:119-141, 1997
24. Faloutsos, C, Equitz, W, Flickner, M, et al: Efficient and effective querying by image content. *J Intell Inf Syst* 3(3/4):231-262, 1994
25. Flickner, M: Query by image and video content: the QBIC system. *Computer* 28(9):23-32, 1995
26. Petrakis, EGM, Orphanoudakis, SC: Methodology for representation, indexing and retrieval of images by content. *Image Vision Comput* 11(8):504-521, 1993
27. Abate, AF, Nappi, M, Tortora, G, Tucci, M: IME: an image management environment with content-based access. *Image Vision Comput* 17:967-980, 1999