Mammographic Image Database (MIDB) and Associated Web-Enabled Software for Research

Mark D. Halling-Brown[1,*], Pádraig T. Looney[2], Mishal N. Patel[3], Lucy M. Warren[2,3], Alistair Mackenzie[2,3], and Kenneth C. Young[2,3]

[1] Scientific Computing, Medical Physics, Royal Surrey County Hospital, Egerton Road, Guildford, GU2 7XX, UK
[2] National Coordinating Centre for the Physics of Mammography, Medical Physics, Royal Surrey County Hospital, Egerton Road, Guildford, GU2 7XX, UK
[3] Department of Physics, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, GU2 7XH, UK

1 Summary

Current efforts relating to the uptake, evaluation and research into digital medical imaging require the large-scale collection of images (both unprocessed and processed) and data. This demand has led us to design and implement a flexible mammographic image repository, which prospectively collects images and data from multiple screening sites throughout the UK. The MIDB has been designed and created to provide a centralised, fully annotated dataset for research purposes. One of the most important features is the inclusion of unprocessed images. In addition to the images and data, systems have been created to allow expert radiologists to annotate the images with interesting clinical features and provide descriptors of these features. MedXViewer (Medical eXtensible Viewer) is an application we have designed to allow workstation-independent, PACS-less viewing and interaction with anonymised medical images (e.g. for observer studies). With these integrated tools, the MIDB has become a valuable resource for running remote observer studies and providing data and statistics for imaging based-research projects. Previously, studies were run by laborious transfers of images to PACS at remote sites and paper-based data manually curated into databases. Apart from the inconvenience, these approaches also suffer from a lack of accurate location information from the paper-based forms.

2 Introduction

There is a need for comprehensive collections of medical images to be made available for research. Among many requirements are the need for unprocessed images, fully annotated cases and fully representative sets from a variety of disciplines and modalities. Collections of images are required to undertake research and development in Computer Aided Detection (CAD), image perception studies, training and quality assurance. The collection of large set of medical images with full annotation for

[*] Corresponding author.

H. Fujita, T. Hara, and C. Muramatsu (Eds.): IWDM 2014, LNCS 8539, pp. 514–519, 2014.

research purposes is challenging. When the gathering of unprocessed images is required, the difficulties increase massively. The need for accurately curated, comprehensive research sets will only increase as the number of new techniques and modalities increases.

Alongside the need for medical images is the need for more rigorous, consistent and timesaving approaches to managing and undertaking image perception studies. Many studies have previously been run with laborious transfers of images to PACS at remote sites and paper-based data manually curated into databases. Apart from the inconvenience, these approaches suffered from a lack of accurate location information from the paper-based data. One of the most notable items that is lacking from many image perception studies is precise cancer location information. In addition, there are many collaborative image-viewing undertakings, which currently require image transfers between PACS and suffer from a lack of ability to centrally annotate cases with descriptions and regions of interests (ROIs). There are many situations where it would be beneficial to be able to have cases reviewed by experts located at remote sites throughout the country, or indeed the world.

The MIDB systems have been designed to be deployed at multiple remote sites. At these sites, it automatically identifies the relevant cases to collect (e.g. screen detected breast cancers) and then obtains the processed and unprocessed images from the local PACS and associated data from relevant local cancer databases. The images and data are then automatically anonymised and transferred to the central storage. In addition to the data, software (MedXViewer) has been created to allow expert radiologists to annotate the images with interesting clinical features and provide descriptors of these features. In order to avoid subsequent manual transfers of images to multiple remote PACS, the images are streamed from the central location.

The MIDB has been created to provide a centralised, fully annotated dataset for research purposes to meet many of the needs outlined above. One of the most important features is the inclusion of unprocessed images. The MIDB has become a valuable resource with integrated tools for running remote observer studies and providing data and statistics for imaging-based research projects. Initially the database was developed as part of a large research project in digital mammography (OPTIMAM). Hence the initial focus has been digital mammography; as a result, much of the work described will focus on this field and all the current images and data are mammographic.

3 Methods

3.1 Image Database and Collection

A semi-automated process for identifying which cases to collect has been developed. Since we are primarily interested in cancer cases we interrogate local databases of patient/case data to identify these among a much larger set of cases. A full description of the processes is described in more detail elsewhere[1].The processes and systems required to allow semi-automated image collection across multiple heterogeneous sites are extremely complex. The database is made up of several relational databases and a file system for storage of the images. In a very simplified manner, the data models can be split into the image database (ImageDB), observer and training study database (StudyDB) and the associated data schemas (AssocDB). The imageDB maintains the DICOM data, the ground-truth and associated data. Comprehensive loading systems have been created which process the new images and insert the appropriate data into the databases. The observer study database holds details of multiple active observer studies, the observers themselves and the marks and progress made in each study (See Fig 1).

Calculated and derived data can be obtained from the images at the time of collection and include various image feature extractions that are useful for classification, CAD and radiomics applications. Additional annotation, such as features identified by Computer Aided Detection (CAD) systems can be inserted into the database at a later date.

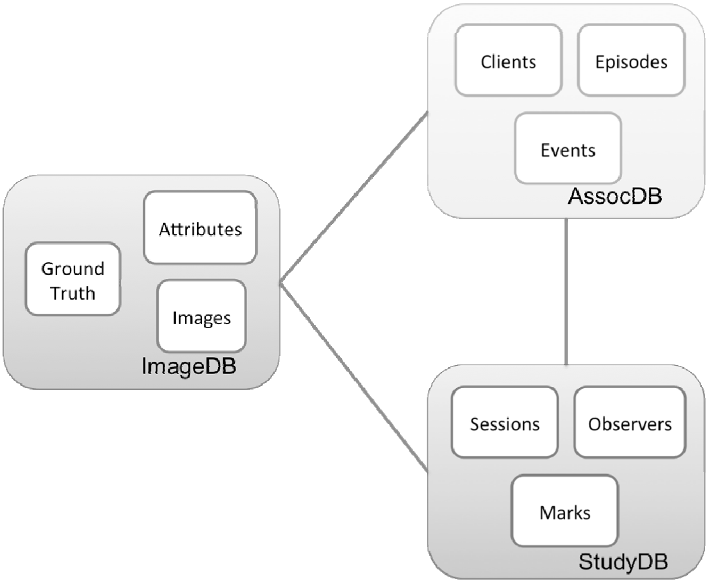

Fig. 1. Overview of main data models involved in OMI-DB. The three main schemas indicated are the Image Database (ImageDB), observer and training study database (StudyDB) and the associated databases (AssocDB).

3.2 Service Layers

In order to allow the tools created to communicate with the OMI-DB, service layer systems were created which manage the messages and transfer between the data and presentation layers. These take the form of web services and are maintained on the same systems as the OMI-DB

3.3 MedXViewer

Various tools enabling interaction with the varying parts of MIDB have been created. Notably, this includes an extensible web-enabled software package designed to facilitate remote studies, collaborative view sessions and training (MedXViewer). Image perception studies and the acquisition of data for research purposes requires the viewing and marking of images, the answering of questions throughout the marking process and the drawing of regions of interests (ROIs) on the images. MedXViewer was initially designed for image perception studies in digital mammography and digital breast tomosynthesis (DBT) but the software can be extended for use in other modalities. One of the key objectives in its development was to allow radiologists to review images and participate in observer studies at multiple remote sites.

MedXViewer was developed in Java and uses the DICOM library Dcm4Che allowing 16-bit images to be created and lookup tables applied. Hence, images can be displayed, as they would appear on a commercial PACS system when MedXViewer is used with clinical quality displays.

The choice of Java as the programming language enables MedXViewer to be used cross-platform on Mac, Linux and Windows environments with no requirement for administrative access or installation. MedXViewer automatically detects the monitor setup and location and places the medical images on the predetermined location. MedXViewer can integrate with the MIDB, allowing images and software to be downloaded from a central store and results to be uploaded to a centralised database. Alternatively the software can be run offline with images and results stored locally. A user is allocated a username and password and their progress and performance can be tracked. The results from MedXViewer can be output in any standard file format (CSV, Excel etc.).

4 Results

The MIDB has been implemented along with the automated collection procedures, anonymisation, associated data gathering, calculated data and expert determined annotations. Currently the full collection system is deployed at three sites and other partial collection processes are in place at two other sites. The database statistics are summarized in Table 1.

When loading the images into the repository, all relevant DICOM tags are extracted to allow a searchable index to be produced. Additionally, CAD predictions are determined and inserted into the database. Expert-determined ground truths are then obtained from our panel of readers (utilising the MedXViewer software) that indicate the relevant ROI and other attributes, such as lesion type and conspicuity. Further annotations are obtained from associated data sources, such as the NBSS. The quantity of additional annotations obtained is large, and includes information on screening history, previous occurrences of cancer, biopsy results and surgical procedures.

Table 1. Database statistics for the MIDB (Correct as of 28/02/2014)

Num of Images	34,014
Num of Studies	4,301
Num of Cases	2,623
Num of Benign Cases	87
Num of Normal Cases	680
Num of Malignant Cases	1,856
Num of Expert Annotated Cases	1,453
Num yet to be Annotated	403

Fig 2 shows the collection of data plotted over time. This is useful to show the projected collection rate and allow us to expect to reach 3,400 cases by the end of April 2014 at the current rate.

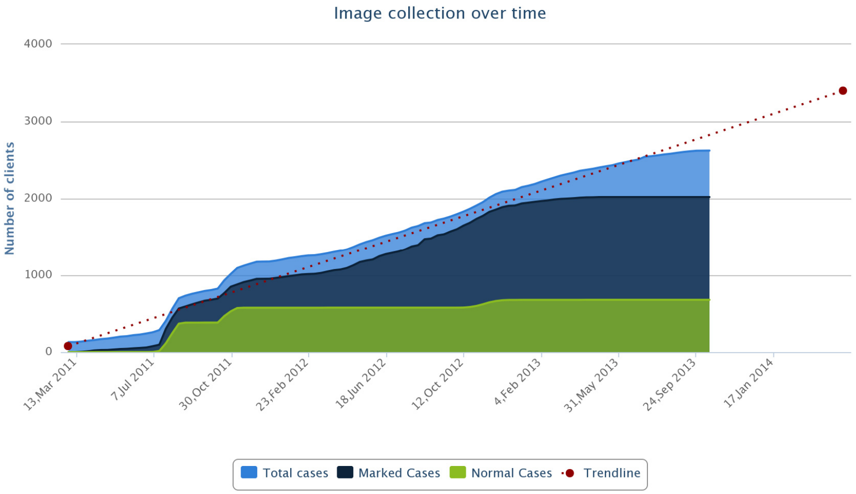

Fig. 2. Image collection over time for the MIDB. Cancer cases often do not become available until several months have passed from the imaging event, hence the lack of recent cases.

Associated tools (including MedXViewer) have been produced which allow interaction with the database in a PACS-less, workstation independent manner. The images from OMI-DB and the associated tools – MedXViewer – have been used in three image perception studies and more are ongoing or planned. The studies investigated the effect of factors such as detector type[2], reader interruption and image processing[3] on cancer detection in digital mammography imaging.

5 Conclusions

A Mammographic Image Database has been created that provides access to a huge number of fully-annotated cases with associated data and unprocessed images. These annotations are obtained from existing cancer databases and from expert opinions. MIDB has already facilitated a number of research projects. The logistics, technology, systems and procedures required bringing together these images and data are extremely difficult to manage and not easily reproduced. We have designed the system to be flexible enough to feasibly allow any site to be added to our collection system. This is facilitated by the provision of multiple differing collection workflows including direct onsite PACS connections. To date we have successfully collected 34,104 2D images from 2,623 clients, run three observer from the MIDB [2,3] and facilitated several studies, including CAD investigations.

In many fields, the retention of unprocessed images is not commonplace. A distinct feature of MIDB is the provision of the unprocessed data that allows certain areas of research to take place, which would otherwise be difficult. An excellent example can be found in Interval Cancer review sessions for mammography. These can be stored, to evaluate whether an abnormality could have been detected on the previous screening films. The provision of unprocessed data from previous screening would allow the comprehensive evaluation of the full raft of factors affecting the reading of images. The effect of image processing can be investigated to find if different processing methods would have facilitated the location of the abnormality.

MedXViewer has been developed which interacts with the MIDB and associated study databases for remote image viewing and interaction. Combined, these provide the ability to run remote paper-less observer studies, provide a training infrastructure or coordinating remote collaborative viewing sessions (e.g. cancer reviews, interesting cases).

Acknowledgements. This work is part of the OPTIMAM project and is supported by CR-UK & EPSRC Cancer Imaging Programme in Surrey, in association with the MRC and Department of Health (England).

References

1. Patel, M.N., Looney, P.T., Young, K.C., Halling-Brown, M.D.: Automated collection of medical images for research from heterogeneous systems: Trials and tribulations. In: Proc. SPIE 9039 Medical Imaging (2014)
2. Mackenzie, A., et al.: Using image simulation to test the effect of detector type on breast cancer detection. In: Proc. SPIE 9037 Medical Imaging (2014)
3. Warren, L.M., Cooke, J., Given-Wilson, R., Wallis, M., Halling-Brown, M.: Effect of image processing on detection of non-calcification cancers in 2D digital mammography imaging. In: Proc SPIE Medical Imaging (2013)