DIGITAL DATABASE FOR SCREENING MAMMOGRAPHY: 1998

M. HEATH[1], K. BOWYER[1], D. KOPANS[2], P. KEGELMEYER JR[3], R. MOORE[2], K. CHANG[1] AND S. MUNISHKUMARAN[1]
[1] *Computer Science and Engineering, Univ of South Florida, Tampa, Florida 33620*
heath, kwb, chang or munish @bigpine.csee.usf.edu
[2] *Department of Radiology, Massachusetts General Hospital 15 Parkman Street, Level 2, Suite 219, Boston, MA 02114*
moore@sisu.mgh.harvard.edu
[3] *Sandia National Labs, Center for Computational Engineering, P.O. Box 969, MS 9214, Livermore, CA, 94551-0969*
wpk@sandia.gov

Abstract. The Digital Database for Screening Mammography[1] is a resource for use by researchers investigating mammogram image analysis. In particular, the resource is focused on the context of image analysis to aid in screening for breast cancer. The database now contains substantial numbers of "normal"and "cancer" cases. This paper describes recent improvements and additions to DDSM.

1. Introduction

Many improvements have been made to the DDSM resource since the 3rd International Workshop on Digital Mammography. These include an enhanced web interface with preview images of each case, additional software and additional cases. This paper describes these improvements and shows some of the less typical cases in the database and image artifacts that may prove difficult for computer analysis methods.

2. Overview of the Database

The Digital Database for Screening Mammography is being constructed to provide the research comminuity with a source of data that can be used to rigorously compare different image analysis techniques. At the time of this paper, the database contains 373 normal and 223 cancer cases. Additional cancer cases and path-proven benign cases are being added as they are processed.

[1]Supported by Department of the Army research grant DAMD17-94-J-4015.

3. The Anatomy of a Case

Each case in the database contains high quality digitized copies of the four images taken in a screening exam (CC and MLO of each breast) as well as ground truth data for cancer and benign cases. Separate directories for each case include the four compressed image files (true lossless JPEG), an ICS file, between zero and four OVERLAY files and a thumbnail mosaic image stored in PGM format.

The ICS file stores the date of study, the patient age, the ACR breast density, the date of digitization, and the size and scanning resolution for each image. A separate OVERLAY file is provided for each image that contains marked lesions. OVERLAY files contain the assessment, subtlety and pathology, and a description and chain code for each lesion. Figure 3 illustrates the contents of an ICS file and the contents of the OVERLAY files (excluding the chain codes) for a case.

4. Issues in the Construction of the Database

4.1. IMAGE QUALITY

Care is taken while clean the films before scanning them. Despite this, there are artifacts in the digitized images. Some artifacts are due to dust or scratches on the film while others are introduced by the scanner. Figure 1 shows extracted regions from images illustrating dust, scratches, and scanner artifacts. Any practical computer assisted screening system that uses digitized films will need to handle these types of artifacts. Even systems using direct digital image capture must cope with pixel non-uniformity and dust in the imaging system.

4.2. GROUND TRUTH

Ground truth is provided by an expert radiologist for each non-normal case. This data is stored in overlay files for each case in a computer readable format. The web site for the database gives detailed descriptions of this data and the format in which it is stored.

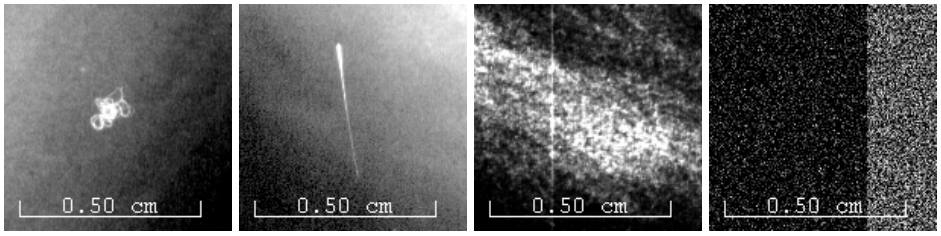


Figure 1. These extracted regions of images have been enhanced to show artifacts they contain. Arranged from left to right, the artifacts are: dust in the camera system, a scratch on the film, a poorly corrected pixel non-uniformity in the scanner (the vertical white line) and under-corrected gain differences between CCD's in the DBA scanner.

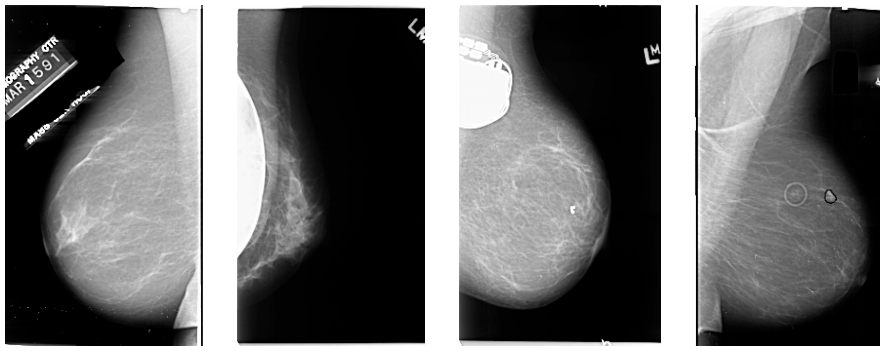


Figure 2. Some cases with unusual attributes. From left to right, a label obscures some breast tissue, a breast implant is visible, a pacemaker is visible and a marking ring is visible. The right image also has a chain code overlayed on it in black indicating a mass.

4.3. ATYPICAL CASES

The cases included in the database were not selected to unnaturally include certain types of lesions, breast densities, patient ages etc. For example, all of the cancer cases found between 1990 through 1996 were candidates for the database. Only cases of low image quality and those that were not accessible for scanning were excluded. As a result, some cases have attributes that may challenge computer analysis techniques for locating lesions. Figure 2 illustrates three atypical cases.

5. Accessing the Database

5.1. WORLD WIDE WEB INTERFACE

The main URL is http://marathon.csee.usf.edu/Mammography/Database.html. This page contains general information about the database including a table of links to previews of every case. Figure 3 illustrates one of the preview pages. We are currently investigating methods for allowing users to search the cases based on the contents of these pages.

5.2. FTP ACCESS

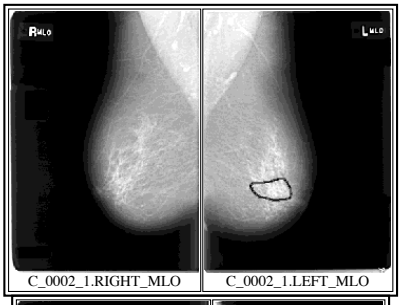
Some of the volumes of cases are available on line by anonymous ftp to ftp://figment.csee.usf.edu/pub/DDSM/cases. These will occasionally be rotated with preference given to new volumes of cases.

5.3. ORDERING TAPES

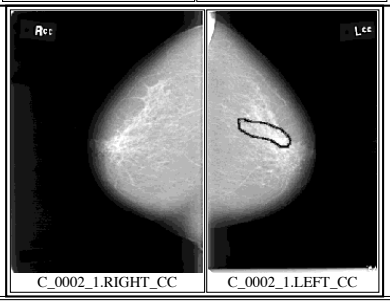
Interested parties can order volumes of cases on 8mm tape. The tapes are TAR archives written on SUN workstation. Typically a tape contains between 3 and 6 Gigabytes of data. The cost is $30 for the first one and $20 for each additional tape. An order form can be obtained from our web site.

Digital Database for Screening Mammography

Volume: cancer_01 Case: C-0002-1



C_0002_1.RIGHT_MLO | C_0002_1.LEFT_MLO



C_0002_1.RIGHT_CC | C_0002_1.LEFT_CC

```
ics_version 1.0
filename C-0002-1
DATE_OF_STUDY 11 5 1992
PATIENT_AGE 72
FILM
FILM_TYPE REGULAR
DENSITY 2
DATE_DIGITIZED 8 18 1997
DIGITIZER LUMISYS LASER
SELECTED
LEFT_CC LINES 5928 PIXELS_PER_LINE 3776 BITS_PER_PIXEL 12 RESOLUTION 50 OVERLAY
LEFT_MLO LINES 5824 PIXELS_PER_LINE 4104 BITS_PER_PIXEL 12 RESOLUTION 50 OVERLAY
RIGHT_CC LINES 5704 PIXELS_PER_LINE 4120 BITS_PER_PIXEL 12 RESOLUTION 50 NON_OVERLAY
RIGHT_MLO LINES 5792 PIXELS_PER_LINE 4144 BITS_PER_PIXEL 12 RESOLUTION 50 NON_OVERLAY
```

```
FILE: C_0002_1.LEFT_MLO.OVERLAY

TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE CALCIFICATION TYPE PLEOMORPHIC DISTRIBUTION SEGMENTAL
ASSESSMENT 4
SUBTLETY 1
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 1
BOUNDARY
```

```
FILE: C_0002_1.LEFT_CC.OVERLAY

TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE CALCIFICATION TYPE PLEOMORPHIC DISTRIBUTION SEGMENTAL
ASSESSMENT 4
SUBTLETY 2
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 1
BOUNDARY
```

Figure 3. The Web page for one case in the database. Each page includes thumbnail images with the chain coded boundaries overlayed on them, the contents of the ICS file and the descriptions in the OVERLAY files. Currently there are 373 normal and 223 cancer cases with pages like this available at http://marathon.csee.usf.edu/Mammography/Database.html.