Cloud-based Medical Image Collection Database with Automated Annotation

Anthony Maeder and Birgit Planitz
School of Computing and Mathematics
University of Western Sydney
Sydney, Australia
a.maeder@uws.edu.au, birgit@planitz.net

***Abstract* - Typical medical image annotation systems use manual annotation or complex proprietary software such as computer-assisted-diagnosis. A more objective approach is required to achieve generalised Content Based Image Retrieval (CBIR) functionality. The Automated Medical Image Collection Annotation (AMICA) toolkit described here addresses this need. A range of content analysis functions are provided to tag images and image regions. The user uploads a DICOM file to an online portal and the software finds and displays images that have similar characteristics. AMICA has been developed to run in the Microsoft cloud environment using the Windows Azure platform, to cater for the storage requirements of typical large medical image databases.**

Keywords - medical imaging; content based image retrieval; cloud computing

I. INTRODUCTION

Large medical image collections provide essential source datasets for research on population health or disease cohort studies, using patient phenotype information derived from the images. For example, we may wish to select cases where brain morphology of a certain kind is conjectured to be related to deposition of plaques linked with the onset of dementia, and finding regions in those images where there is a higher expectation of plaque formation. Presently, *annotation of medical images* for such work is conducted either manually by highly specialist expert viewers or radiologists [1], or by complex proprietary software such as computer-assisted-diagnosis. This is a time consuming process, subject to error and bias. Selection of images according to established Content Based Image Retrieval (CBIR) criteria such as "similar to a given image" or "containing range of characteristics" would be far more objective if it could be conducted automatically. To achieve this for medical images requires a set of functions which cater for the typical components of image similarity, and can be tuned to suit different medical image types.

The *Automated Medical Image Collection Annotation (AMICA)* toolkit for implementing the above CBIR criteria has been developed to address this need. It is intended to contribute to advances in clinical research by permitting wider use of medical images than is currently practiced, such as:

a) Population health and epidemiological studies that rely on content analysis of images and related patient information will be much easier to conduct via an electronic medium than by using expert human readers, and availability of compatible computer processible image data widely from various sources will allow a much fuller analysis which truly covers most of the population;

b) Cohort studies for cases with particular physiological or phenotypical profiles will be able to source and include enough cases to provide high statistical power, allowing more individualised risk factors to be assessed and thus allowing screening and staging processes to be optimised. Cases will also be selectable on a wider basis than is achievable now from patient information in electronic records, by use of image content analysis;

c) Education and training/credentialing of radiographers, radiologists and other clinicians who are involved with image interpretation will be more effective because it will be possible to select instances of images which demonstrate particular visual aspects, or correspond to types of cases where reading performance improvement is desirable for that individual.

The AMICA software is appropriate for the above situations because both the medical image database and automated annotation tools are stored in a cloud environment. That is, we make use of the flexibility and scalability of the cloud to grow our application to suit researchers' needs. We envisage that our application will grow beyond its initial pilot implementation to a wide ranging application suitable for a range of modalities and anatomical regions.

This paper provides a brief review of existing medical image collections and annotation tools, and describes how AMICA differs from other software. We discuss how Microsoft technologies have been used to deploy our web application in the cloud. We also detail the CBIR functions used for image annotation; our CBIR algorithm works on the principle of "find other images like my given image". We conclude with a discussion on leveraging the cloud environment to expand our pilot implementation of AMICA.

II. RELATED RESEARCH

Many widely used digital medical image collections have previously been established but these are generally used as raw

 ISBN: 978-1-61208-185-4

data source only, without related toolsets. Providing associated functionality to allow specific types of operations to be performed on these images has proved beneficial in some cases (e.g., brain image registration [2]; brain atlases [3]). However, toolset development for image analysis functions on medical images has tended to be ad hoc, with Open Source options proliferating.

Several major organisations, particularly in the U.S., such as the National Cancer Institute, have made medical images databases and associated search tools available to the wider research community. A popular project is the Visible Human, which is a database of transverse CT, MR and cryosection images of three dimensional (3D) anatomical representations of male and female cadavers [4]. The National Institute of Health has made these datasets available for study, and also developed ITK, an open source toolkit for image registration and segmentation [5]. The Biomedical Informatics Research Network (BIRN) has made a list of tools available for data storage and medical image processing [6], and has also developed a downloadable Human Imaging Database [7]. The National Cancer Institute developed Annotation Imaging Markup, a downloadable manual annotation tool that enables easy and automated image searching [8]. The Institute also produced the National Biomedical Imaging Archive, which includes a manual annotation option and a web portal for accessing medical images [9]. Other large medical image databases (e.g., the Singapore National Medical Image Resource Centre [10]) are searchable via keywords but not image content.

CBIR has been applied to medical image datasets, however this has generally been limited to specific cases. For example, tumour detection [11], retrieval of lung images [12], 3D MR images [13], 3D MR and CT images [14], or image shape for retrieving pathology [15]. In a project with broader scope, Principal Component Analysis (PCA) has been used for medical image classification [16]. In a system similar to ours, PCA was applied to find closest matches between features of candidate brain images and a set of test images. However, the system is a form-type application, rather than web-based, therefore limiting its capacity to a local setting. Also, results were limited to brain datasets. Mojsilovic and Gomes developed a web-based system for classifying numerous image datasets according to modality [17]. Their program searched large medical image databases and performed the classification automatically. The objective of the modality classification was to categorise images so that domain-specific CBIR functions could later be applied to datasets in each modality. The approach is flexible in that it encompasses existing online databases without users having to load their own. However, this system becomes superfluous in the case of DICOM data, where modality information is stored in the data file.

It would appear that existing medical image repositories and retrieval systems were typically designed with specific applications in mind. Alternatively, our AMICA software provides a simple but comprehensive toolset that could be established as a baseline. Our software takes an image and finds similar images according to some basic CBIR criteria. We believe that this approach, coupled with our leveraging of Microsoft's cloud environment makes for a flexible and scalable tool for a wide variety of medical images.

III. CLOUD-BASED AUTOMATED ANNOTATION TOOLS

This project uses the Windows Azure Platform for data management by exploiting the cloud model. By hosting the AMICA software in the cloud, we tap into the potential of growing the application to use significant volumes of medical images. Individually, medical images tend to be of large sizes (e.g., >25MB for a mammogram), so a substantial collection quickly causes an explosion in server storage requirements. Our objective is to make our database scalable and to provide sufficient storage space for users to upload the datasets that they wish to annotate automatically. Using the cloud, we have the benefit of flexibility, because we can scale up our storage needs as required. We also run our web interface and worker role (for CBIR) in the cloud, where they are stored multiply to ensure that the web interface is constantly available.

Our software consists of three components:

1. Storage Table (SQL Azure)
2. CBIR (Azure Worker Role)
3. Website (Azure Web Role)

We have set up the application using .net and C# in the Visual Studio environment, which extend naturally when deploying an application to Azure Platform.

A. Storage Table

Our first important design choice was whether to store medical image data within Azure's storage environment, or to use a SQL Azure table. We consulted the MSDN developer magazine, which advises [18]:

"If you have an application that requires data processing over large data sets, then SQL Azure is a good choice. If you have an app that stores and retrieves (scans/filters) large datasets but does not require data processing, then Windows Azure Table Storage is a superior choice."

Hence, we have used a SQL Azure table because our project involves significant image processing on large numbers of medical images.

We designed our data storage such that DICOM files and their associated image attributes, which are retrieved using CBIR, are stored as rows in a table, as shown in Table I.

Each entity (table entry) has a unique ID, which is generated automatically and consists of DICOM header elements.

We list Modality and Anatomical Region for each DICOM file explicitly. This reduces our initial search function, i.e., we assume that users wish to find images similar to their given image, which is of a specific body part that was captured using a specific modality. In terms of image processing, it also makes sense to segment the database in this manner, as different image processing functions apply to different image types.

 ISBN: 978-1-61208-185-4

TABLE I. MEDICAL IMAGE DATABASE TABLE

Column Name	Data Type
ID	int
Modality	nvarchar
AnatomicalRegion	nvarchar
DicomInfo	text
DicomImage	image
ImageAttributes	text

DICOM files are separated into information and image files. The aim of separating the files is also to speed up the search function. For example, a user may be studying dementia, and thus may want to only find brain images (that are like their given image) for patients of a specific age. In a future implementation, we will apply a pre-filter to scan DICOM information and retrieve relevant DICOM files, before performing more time-consuming image processing.

DICOM Images are stored as image data types, which are data types that hold any type of binary data. We read BLOBs (Binary Large Objects) in as streams and manipulate/display images according to the information (e.g., Bit Depth) extracted from the DICOM Info file.

The ImageAttributes text file is the file produced by the CBIR function. Image features and their values (e.g., foreground/background intensity threshold) are stored so that they can be compared to a given image in the retrieval process.

B. CBIR

The CBIR toolset applies specific image analysis tasks on medical digital images, providing information on fundamental visual appearance characteristics of the images for metadata annotation purposes. The tools support two different types of annotations: (i) overall image characteristics and (ii) location and extent of specific features (i.e., regions of interest). The provision of these annotations allows matching to be performed between pairs of candidate images, to extract groups of images within a prescribed similarity envelope. For example, given a sample mammogram indicative of a breast with dense texture, other images with a comparable texture density can be extracted by comparison of the breast texture annotation values. We have not fully defined the range of ideal or conventional annotations as this aspect is work in progress.

Typically, in medical research, image subsets are extracted from image collections based on the appearance to human observers of image visual content characteristics such as statistical properties (e.g., textures) or structural features (e.g., regions of interest). For an automated annotation process, application of a sequence of software tools is required to allow users to undertake the following tasks on a given image collection:

(i) Define foreground (i.e., zone of relevance or region of interest) versus background (i.e., zone to ignore) image components;

(ii) Select and apply parameters for characterizing overall image properties (e.g., statistical texture/intensity profiles);

(iii) Select and apply parameters for characterizing region of interest features (e.g., statistical density/edge properties);

(iv) Select and display images having candidate annotations for properties/features within specified ranges;

(v) Iterate in repeated cycle of reapplying the above steps with parameter variations to refine the results.

A major advantage of our approach is that these types of annotation helps avoid the need to reapply image analysis computations each time the image collection is searched: it is analogous to pre-scanning a text data file for keywords and saving these as indexed tags for the file to be available to search tools. Allowing image subsets to be extracted rapidly and consistently using these annotations will improve speed and ease of use in research projects that rely on them.

The CBIR toolset is being developed as a Worker Role in Windows Azure. Images are dispatched to a queue for background processing [19]. The worker role takes images from the queue and processes them; it generates thumbnails for each image in our current pilot implementation.

C. Website

The AMICA website has functions for:

(i) Uploading DICOM files (MR Brain and Mammograms in the current implementation, as described below); and

(ii) An output panel that displays (as thumbnails) the user's loaded image on the left and similar images, as found by CBIR, as on the right hand side (see Fig. 1).

We are tuning and testing our pilot implementation by using two distinct medical image types.

We use MR brain images as these constitute one of the most prolific forms of medical image data present across the health sector. Datasets, such as the Vasari collection are downloadable and are used to test image annotation algorithms, specifically, "to validate the use of medical images as predictive biomarkers for cancer diagnosis" [20]. These DICOM files are suitable for testing Cloud based web and worker applications because the images are small and thus enable efficient testing. We also test with a Matlab brain MR dataset for the same reason [21].

We are also testing and tuning with mammograms because large scale collections exist de facto in countries where there is a national breast cancer screening program. Mammogram image collections were first developed using scanned films in 1980s by MIAS, and subsequently USF, which has become the benchmark data set with several thousand images. Some proprietary collections exist for commercial research by CAD and imaging manufacturers (e.g., SECTRA, R2). The possibility exists that our AMICA software can be used for data management and annotation such that national coordination of breast screening images (e.g., as proposed in the UK e-Diamond project [22]) could ensue.

 ISBN: 978-1-61208-185-4

In Australia, the wide establishment and stimulation of eResearch infrastructure offers an opportunity contribute to emerging platforms and tools for basic data repository functions, and to enhance these by developing some more sophisticated ones suited to this application area. Most eResearch tools developed locally have been concerned with text or symbolic data annotation and analysis, so our software may contribute comparatively novel tools by addressing the digital image space.

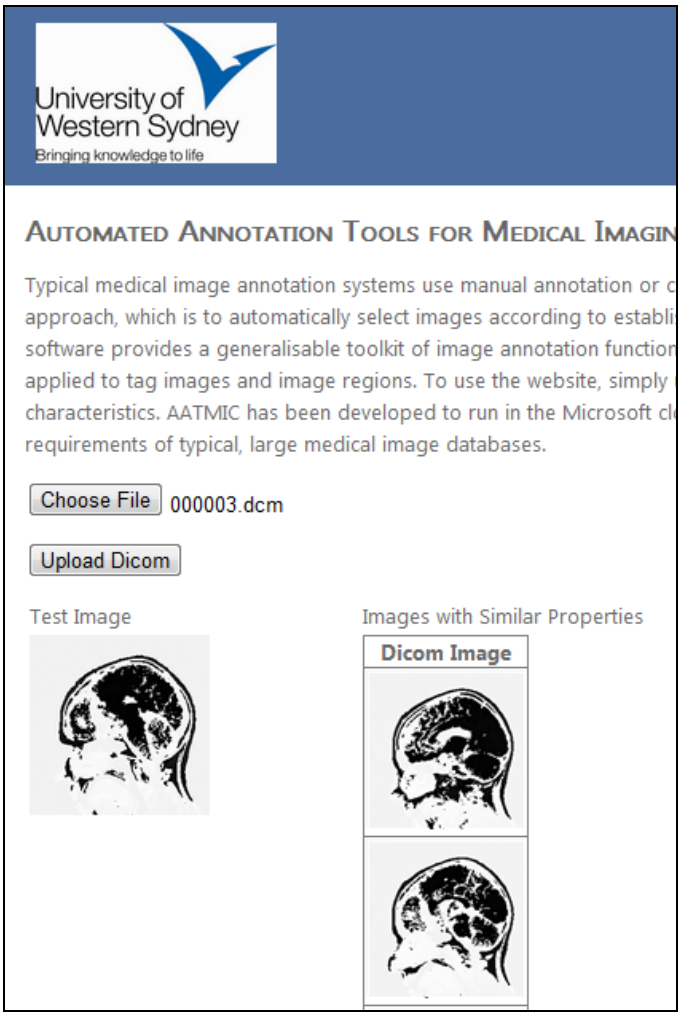

Figure 1. Partial screenshot of AATMIC Website

Like our SQL table and CBIR worker role, the AMICA web interface also operates from Windows Azure. Although the storage requirements for the online portal are not overwhelming (we display thumbnails for all but individually selected images), we make use of Azure's triplicate storage of the website. That is, the redundancy ensures that the website is always up and running should one, or even two, servers be down [24].

IV. DISCUSSION

Our short term goal is to complete the pilot implementation of the AMICA software. Then, we will extend the implementation as follows:

(1) The final pilot implementation will be tuned and tested using a greater number of Vasari MR brain datasets and our entire collection of mammograms (approx. 500 x 50MB image sets). Further testing will therefore cover a much wider range of different image content variations. These results will be used to evaluate computational efficiency of the tools and success of the automated annotations, as rated by comparison with human expert opinions.

(2) We will then extend our CBIR toolset to include a wider range of modalities and anatomical regions (e.g., CT brain, CT chest, CR chest).

(3) To cope with the predicted increase in data, we will provide a separate form for pre-filtering DICOM files. As mentioned earlier, before any image processing takes place, DICOM Information files could be pre-filtered so that only a relevant subset of images is returned for a given search. The pre-filtering function should be turned off if the user wishes to find test images in the entire database for a given Modality and Anatomical Region. We have not yet implemented this function, but see the need to do so as the medical image database grows and as searches become more specific (e.g., the dementia project mentioned earlier).

(4) In terms of CBIR, further benefits could be obtained by selection of similar cases to that in a given image, and selection of archetypical images for certain image features, using feature vectors and classification methods on certain combinations of annotation fields. This would contribute to fuller understanding of how such combinations occur, as well as providing suitable sample images for clinical education and training purposes.

V. CONCLUSION

This paper has described an ongoing project in automated medical image annotation for management of image collections. The software that is currently under development addresses problems hampering existing systems, which either use manual annotation or complex proprietary software. Alternatively, we have designed the AMICA software to automatically select images according to established Content Based Image Retrieval (CBIR) criteria. Images are tagged and managed via a generalisable toolkit of image annotation functions. AMICA's website, storage and worker roles are being developed to run in the Microsoft cloud environment, to exploit the storage requirements of typical large medical image databases. Also, we ensure that the website is always accessible to users via Azure's triple redundancy storage system. The approach could be generalized to use other cloud environments, provided they offer similar functionality.

ACKNOWLEDGMENT

The support of Microsoft Research in funding this project is gratefully acknowledged.

 ISBN: 978-1-61208-185-4

REFERENCES

[1] C. E. Chronaki, X. Zabulis, and S. C. Orphanoudakis, "I2net medical image annotation service," *Informatics for Health and Social Care*, vol. 22, pp. 337-347, 1997.

[2] J. M. F. Jay West, et al., "Comparison and evaluation of retrospective intermodality brain image registration techniques," *Journal of Computer Assisted Tomography*, vol. 21, pp. 554-566, 1997.

[3] Wieslaw L. Nowinski, et al., "Multiple Brain Atlas Database and Atlas-Based Neuroimaging System," *Computer Aided Surgery*, vol. 2, pp. 42-66, 1997.

[4] National Library of Medicine. *The Visible Human Project* [website]. Available: http://www.nlm.nih.gov/research/visible/visible_human.html (accessed: Nov 2011)

[5] National Library of Medicine. *ITK Insight Toolkit* [website]. Available: http://www.itk.org/ (accessed: Nov 2011)

[6] BIRN. *Tools*. [website]. Available: http://www.birncommunity.org/resources/tools/ (accessed: Nov 2011)

[7] NITRC. *Human Imaging Database (HID)*. [website]. Available: http://www.nitrc.org/projects/hid/ (accessed: Nov 2011)

[8] National Cancer Institute. *Annotation Imaging Markup (AIM)*. [website]. Available: https://wiki.nci.nih.gov/display/AIM/Annotation+Imaging+Markup+%28AIM%29 (accessed: Jan 2012)

[9] National Cancer Institute. *National Biomedical Imaging Archive*. [website] Available: https://wiki.nci.nih.gov/display/ImagingKC/Imaging+Knowledge+Center (accessed: Jan 2012)

[10] Guo-Liang Yang and C. T. Lim, "Singapore National Medical Image Resource Centre (SN.MIRC): A World Wide Web Resource for Radiology Education," *Annals Academy of Medicine*, vol. 35, pp. 558-563, 2006.

[11] Philip Korn, Nicholas Sidiropoulos, Christos Faloutsos, Eliot Siegel, and Z. Protopapas, "Fast and Effective Retrieval of Medical Tumor Shapes," *IEEE Trans. on Knowledge and Data Engineering*, vol. 10, pp. 889-904, 1998.

[12] C. E. B. Chi-Ren Shyu, Avinash C. Kak, Akio Kosaka, "ASSERT: A Physician-in-the-Loop Content-Based Retrieval System for HRCT Image Databases," *Computer Vision and Image Understanding*, vol. 75, pp. 111-132, 1999.

[13] G. S. Jérôme Declerck, Jean-Philippe Thirion, Nicholas Ayache, "Automatic retrieval of anatomical structures in 3D medical images," *LNCS: Computer Vision, Virtual Reality and Robotics in Medicine*, vol. 905, pp. 151-162, 1995.

[14] Yanxi Liu, Frank Dellaert, and W. E. Rothfus, "Classification Driven Semantic Based Medical Image Indexing and Retrieval," Robotics Institute, Carnegie Mellon University, Technical Report CMU-RI-TR-98-25, 1998.

[15] Dorin Comaniciu, David Foran, and P. Meer, "Shape-based image indexing and retrieval for diagnostic pathology," Proceedings of the 14th International Conference on Pattern Recognition, pp. 902-904, 1998.

[16] Usha Sinha and H. Kangarloo, "Principal Component Analysis for Content-based Image Retrieval," *RadioGraphics*, vol. 22, pp. 1271-1289, 2002.

[17] A. M. a. J. Gomes, "Semantic based categorization, browsing and retrieval in medical image databases," Proceedings of the 2002 International Conference on Image Processing, pp. III-145 - III-148, 2002.

[18] J. Fultz. SQL Azure and Windows Azure Table Storage. *MSDN Magazine*. Available: http://msdn.microsoft.com/en-us/magazine/gg309178.aspx (accessed: Nov 2011)

[19] Microsoft Corp. *Windows Azure Platform Training Course*. [website]. Available: http://msdn.microsoft.com/en-us/windowsazure/wazplatformtrainingcourse.aspx (accessed: Nov 2011)

[20] National Cancer Institute. *VASARI*. [website]. Available: http://cabig.cancer.gov/action/collaborations/vasari/ (accessed: Jan 2012)

[21] J. Mather. *DICOM Example Files*. [website]. Available: http://www.mathworks.com/matlabcentral/fileexchange/2762-dicom-example-files (accessed: Nov 2011)

[22] Michael Brady, David Gavaghan, Andrew Simpson, Miguel Mulet Parada, and R. Highnam, "eDiamond: a Grid-enabled federated database of annotated mammograms," in *Grid Computing – Making the Global Infrastructure a Reality*, Fran Berman, G. Fox, and T. Hey, Eds., Chichester: John Wiley & Sons, pp. 923-944, 2003.

[23] N. B. C. Foundation. *Lifepool*. [website]. Available: http://www.lifepool.org/ (accessed: Nov 2011)

[24] Charlie Kaufman and R. Venkatapathy, "Windows Azure™ Security Overview," 2010.

 ISBN: 978-1-61208-185-4