Building a library of annotated pulmonary CT cases for diagnostic aid

Adrien Depeursinge[a], Henning Müller[a], Asmaa Hidki[a], Pierre-Alexandre Poletti[b], Thierry Rochat[c] and Antoine Geissbuhler[a]

[a] *Service of Medical Informatics, University and Hospitals of Geneva (HUG), Switzerland*

[b] *Emergency Radiology, HUG, Switzerland*

[c] *Pulmonary Department, HUG, Switzerland*

Abstract.

Interstitial lung diseases (ILDs) are characterized by diverse disorders of the lung tissue with frequently confusing symptoms. The first imaging method used for diagnostics is the chest x-ray but it is not always explicit enough. High-resolution computed tomography (HRCT) of the chest contains essential visual data for the characterization of ILDs. The interpretation of HRCTs is difficult even for specialists as many diseases are rare. An image-based diagnostic aid tool could bring precious elements to less experienced radiologists or non-chest experts. It can thus replace the search in printed reference books.

The development of these tools requires creating a library of pulmonary CT cases, which incorporates annotations of regions of abnormal lung tissues. This paper details the steps for building a database of ILD cases and an annotation tool. The precision of the annotations is fundamental for the accuracy of the diagnostic aid tool. The annotation software is implemented in a web-based manner, allowing high-quality creation of regions of interest (ROIs) in any layer of a CT volume. Finally, a correct interpretation of HRCTs requires metadata of the concerned case integrated in the database.

Currently, only the annotation tool and the metadata definition are available but cases are identified and will over time populate the database.

Keywords: high-resolution lung CT, diagnostic aid, computer-assisted image analysis, classification, texture analysis.

1. Introduction

Interstitial lung diseases (ILDs) are a relatively heterogeneous group of around 150 different diseases of the lung tissue with often very unspecific symptoms. Many of the diseases are rare and their differential diagnostics is regarded as difficult. The most common imaging procedure used for diagnostics is the chest x-ray, because of its low costs and weak radiation exposure. However, chest x-rays are negative in a large portion of diseases and often unspecific [1]. High-resolution computed tomography (HRCT) of the chest often provides more information for diagnostics of interstitial lung diseases. Nevertheless, a correct interpretation of these images is difficult even for experienced chest radiologists and lung specialists. In order to share the expertise of these specialists for teaching and diagnostic assistance we plan to create a digital library of annotated lung CT series with specialist annotations. Recent studies [2,3] show that diagnostic aid systems can improve the diagnostic quality significantly, especially when the radiologists are less experienced.

In [4], a diagnostic aid framework based on texture analysis is proposed. Such a tool has to be built on a large and well-annotated database. This paper details the creation of a database of annotated HRCT. Chest radiologists and lung specialists will annotate relevant HRCT of pathologic or non-pathologic cases.

The labelled regions of interest (ROI) corresponding to the various interstitial lung diseases are associated to each HRCT volume.

The framework study showed that annotating each slice of the 3D chest stack separately is not comfortable for the radiologist. Thus, in order to improve the quality of the annotations in the database, new software allowing three-dimensional delineation with a convenient user interface is implemented. The list of ROIs corresponding to the annotations carried out by the radiologist is stored in an XML file in parallel with the original DICOM file containing the image stack. Standard formats exist for storing ROIs in an XML file, which allows the edition of ROIs from other DICOM viewers such as Osirix [5].

Compared to other content-based image retrieval systems for interstitial lung diseases [2], our database will be built from diagnoses carried out by chest radiologists in collaboration with lung specialists. Indeed, the diagnostic process is driven as follows: HRCTs are first analysed by chest radiologists and suspicious areas are communicated to a lung specialist, who, together with other clinical data of the patient, provides a final diagnosis or asks for additional examinations.

Creating a library of HRCTs from clinical data sets generates privacy problems. An agreement of the ethics commission was obtained in February 2006 in order to begin the data collection and annotation retrospectively. Cases of typical lung diseases from clinical routine are currently being collected for inclusion into the database.

2. Database creation steps

The different steps for creating a large reference database of annotated pulmonary CT cases with high-quality annotation are described in this section. The database is created specifically to integrate the data required for the development of a content-based lung image retrieval system.

2.1.annotation software

The creation of a specific library of annotated pulmonary CT cases for diagnostic aid implies the development of adequate software for the annotation of the relevant ROIs in images. Since the annotation tool will be used by radiologists and lung specialists, it has to work in the hospital environment and has to be accessible from different stations working under various operating systems. In addition, the framework study [4] showed the need of performing annotation on the entire chest image stack to have a global view of the native three-dimensional data. These constraints direct us towards a web-based Java development. In order to inherit from an entire set of drawing tools and from management of image stacks and image formats, the annotating software is implemented as an ImageJ [10] plugin. ImageJ is an open-source image processing platform developed in Java, which allows the development of customized plugins through a collection of image management classes and routines.

2.2.Full DICOM format

Up to now, many tools for diagnosis assistance of lund CTs such as ASSERT [6] are based on 8-bit jpeg images, which allow only 256 grey levels instead of the 10-14 bit output of modern scanners stored in DICOM format. This reduction requires an irreversible level/windowing operation on the initial Hounsfield scale. Such a reduction is annoying since common interstitial lung diseases are represented in the entire Hounsfield range. For example, emphysema is characterized by the destruction of pulmonary tissue and thus represented by “air” corresponding to values from -1000 to -100 Hounsfield Units (HU). On the other hand, calcified areas are corresponding to values around 300 HU. Thus, even if a mapping on 256 grey levels is mandatory for displaying images on screen, the possibility to access original data laid out in the whole Houndsfield range is necessary for an efficient classification of the various interstitial lung diseases.

In addition, as explained in section 2.1, a global view of the three-dimensional chest image is required for high-quality delineation of the ROIs, which finally drove the choice of image format towards full DICOM files. In parallel, the labelled ROIs are stored in a XML file according to a standard format, which allows editing ROIs from other DICOM viewers such as Osirix [5].

2.3.Annotations for classification

In order to retrieve similar cases from the image library, the submitted image data has to be classified into

the diverse diseases. This classification will be based mainly on texture features, which characterise the properties of the lung tissue well [4]. The classifier for the cases first needs to be trained. High quality annotations are directly related to the accuracy of the output. The database needs to include healthy cases as well as pathologic cases with the same slice thickness and slice distance. It is important to note that prototypically healthy regions have to be annotated by the radiologist as well so that a classifier can get a good idea of the texture of healthy tissue. Indeed, the first step in the diagnostic process is to find out whether the tissue is abnormal or not.

The framework study [4] showed that annotations dedicated to machine learning have to be done in a painstaking fashion. One typical problem is that manually marked ROIs are too large around the pathologic tissue. Erroneous results with a part of the healthy tissue also being marked can be seen in Figure 1.

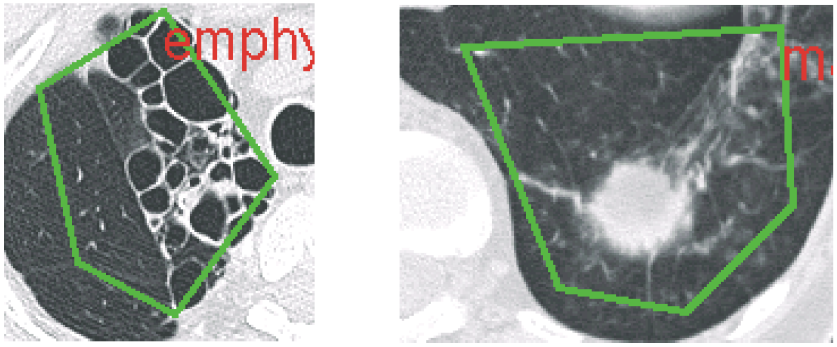

Figure 1: Annotations intended for machine learning have to be carried out with precautions; regions of interest that are not well marked will result in classification errors (emphysema marked on the left and a macronodule on the right).

2.4.Additional data required

Final diagnostics for interstitial lung diseases cannot be carried out exclusively from the visual content of images. In clinics, the image description of the radiologist is also only part of the picture and the lung specialist finally determines the diagnosis. In order to retrieve similar cases, the set of clinical data such as age, weight, and environmental exposures are as significant as visual content of image datasets [9]. Content-based access to data by visual features is complementary to text-based queries and is unlikely to ever replace them completely [11].

3. Results

A web-based graphical tool was implemented in Java based on the open source program ImageJ [10] to allow for high-quality annotation of the regions of interest (ROIs) in the CT series. From the program is based on Java and thus usable from any terminal PC in the hospital without the need to install additional programs, which is often not possible as the users are not administrators of their machine for security reasons. The use of Java also makes the system platform independent. The radiologist opens an entire DICOM series at once and then draws precise ROIs in any layer of the CT volume. Different drawing and visualization tools are available in the ImageJ toolbar. In addition, ImageJ contains several routines for managing DICOM files and allows basic three-dimensional representations. ROIs are stored in XML files by storing the coordinates of each point on the boundary. This lets us reuse the ROIs easily in a variety of contexts, and also by other program such as OsiriX [5].

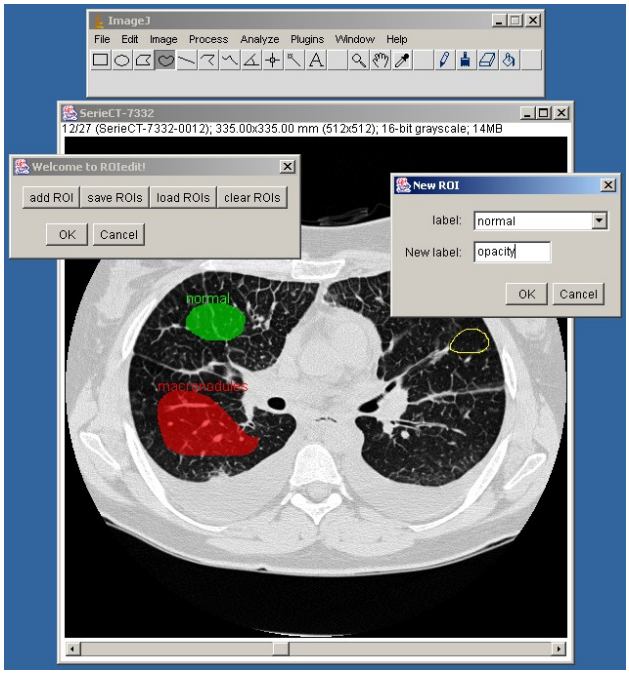

Figure 2: A screenshot of the graphical tool for the annotation of image regions.The MD can easily browse the whole stack of the three-dimensional image using the horizontal scrollbar located at the bottom of the image.

With our annotating tool and the work of the chest radiologists and lung specialists, we expect to collect at least 100-200 series of 50-60 HRCT images for the new database. This will allow for a good representation of each class corresponding to the various interstitial lung diseases for classification.

4. Discussion and Conclusions

At the time of the pilot study [4], a strategy allowing acquiring knowledge and experience from chest radiologists was set up. Today, we have a mature approach inherited from past experience. Indeed, the new database will be constituted of diagnoses worked out by chest experts. The system will take into account clinical parameters extracted from the electronic patient record. Thus, the database will contain as well annotated HRCT dataset as well as clinical data from the patient file.

The library will constitute a solid basis for the development of an innovative teaching and diagnostic aid tool. The content-based image retrieval system will provide a series of similar cases with patient history, and diagnoses when submitting a case currently under observation. This multimodal retrieval will take into account visual features and structured metadata on the patient.

5. Future Work

Once a large database of HRCT series and metadata is established, a strategy has to be set up in order to index its content for the retrieval of similar cases. This retrieval strategy will be structured as follows.

Research is planned for the selection of the visual features required and on the weighting of visual as well as non-visual features. To retrieve visual information from HRCT images, texture analysis of lung tissue and segmentation of the lung based on mathematical morphology will be used [4,7,8]. Additionally, clinical data from the electronic patient record will constitute features for retrieval.

The selected features will be used for classifier training based on the annotated ROIs. Then, features from the unknown case are extracted and submitted for classification. The pilot study showed that classifiers based on Support Vector Machines (SVM) often give best overall results [4].

6. Acknowledgements

This work was supported by the Swiss National Science Foundation (FNS) with grant 205321-109304/1, the equalization fund of University and Hospitals of Geneva (grant 05-9-II) and a grant from the Swiss Confederation for the work of Asmaa Hidki.

7. References

[1] P Stark, High resolution computed tomography of the lungs, Uptodate, February 7, 2005.
[2] AM Aisen, LS Broderick, H Winer-Muram, CE Brodley, AC Kak, C Pavlopoulou, J Dy, CR Shyu, A Marchiori, Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment, Radiology, 228, pp. 265-270, 2003.
[3] H Müller, N Michoux, D Bandon, A Geissbuhler, A review of content-based image retrieval systems in medicine – clinical benefits and future directions, International Journal of Medical Informatics, 73, pp 1-23, 2004.
[4] H Müller, S Marquis, G Cohen, PA Poletti, C Lovis, A Geissbuhler, Automatic abnormal region detection in lung CT images for visual retrieval, *Swiss Medical Informatics*, 2006 - to appear.
[5] A Rosset, L Spadola, O Ratib, OsiriX: An open-source software for navigating in multidimensional DICOM images, Journal of Digital Imaging, Vol 17, No 3, pp.205-216, 2004.
[6] CR Shyu, C Brodley, A Kak, A Kosaka, ASSERT: A Physician-in-the-loop content-based retrieval system for HRCT image databases, Computer Vision and Image Understanding 75 (1-2), pp. 111-132, 1999.
[7] R Uppaluri, EA Hoffman, M Sonka, GW Hunninghake, G McLennan, Interstitial lung disease: A quantitative study using the adaptive multiple feature method, Am J Respir Crit Care Med, Vol 159, pp 519-525, 1999.
[8] EA Hoffman, J Reinhardt, M Sonka, BA Simon, J Guo, O Saba, D Chon, S Samrah, H Shikata, J Tschirren, K Palagyi, KC Beck, G McLennan, Characterization of the interstitial lung diseases via density-based and texture-based analysis of computed tomography images of lung structure and function, Academic Radiology, 10:1104-18, 2003.
[9] A Hidkii, H Müller, A Depeursinge, PA Poletti, A Geissbuhler, Putting the image into perspective: the need for domain knowledge when performing image-based diagnostic aid, 2006.
[10] http://rsb.info.nih.gov/ij/
[11] http://www.sim.hcuge.ch/medgift/