Computerized Medical Imaging and Graphics 36 (2012) 227–238

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

# Building a reference multimedia database for interstitial lung diseases

Adrien Depeursinge[a,b,*], Alejandro Vargas[b], Alexandra Platon[c], Antoine Geissbuhler[b], Pierre-Alexandre Poletti[c], Henning Müller[a,b]

[a] University of Applied Sciences Western Switzerland (HES-SO), TechnoArk 3, CH-3960 Sierre, Switzerland
[b] University and University Hospitals of Geneva, Service of Medical Informatics, 4, rue Gabrielle-Perret-Gentil, CH-1211 Geneva 14, Switzerland
[c] University and University Hospitals of Geneva, Service of Emergency Radiology, 4, rue Gabrielle-Perret-Gentil, CH-1211 Geneva 14, Switzerland

ARTICLE INFO

*Article history:*
Received 9 November 2010
Received in revised form 13 June 2011
Accepted 6 July 2011

*Keywords:*
Interstitial lung diseases
Multimedia database
High-resolution computed tomography
Computer-aided diagnosis
Content-based image retrieval
Case-based retrieval

ABSTRACT

This paper describes the methodology used to create a multimedia collection of cases with interstitial lung diseases (ILDs) at the University Hospitals of Geneva. The dataset contains high-resolution computed tomography (HRCT) image series with three-dimensional annotated regions of pathological lung tissue along with clinical parameters from patients with pathologically proven diagnoses of ILDs. The motivations for this work is to palliate the lack of publicly available collections of ILD cases to serve as a basis for the development and evaluation of image-based computerized diagnostic aid. After 38 months of data collection, the library contains 128 patients affected with one of the 13 histological diagnoses of ILDs, 108 image series with more than 41 l of annotated lung tissue patterns as well as a comprehensive set of 99 clinical parameters related to ILDs. The database is available for research on request and after signature of a license agreement.



## 1. Introduction

In order to understand the underlying mechanisms of groups of diseases, the first requirement is to collect a sufficient number of cases that are representative of the various realizations of the studied diseases. Moreover, the quality of the acquired data is a key for carrying out non-biased evaluations of computer-aided diagnosis (CAD) systems. With the aim to enhance the disease therapy in clinical routine, the library of cases has to be as representative as possible of the hospital's population, meaning that the cases have to be chosen randomly among the entire population if possible.

These requirements are of high importance when building image-based computerized diagnostic aid based on medical image processing [1,2]. A high-quality multimedia collection of cases containing annotated image series and associated clinical parameters is required to ensure the success of a CAD system at the time it will be integrated into clinical routine [3].

On the one hand, the database constitutes a basis for developing computerized tools [4,5] such as automatic detection of abnormal pulmonary tissue types in high-resolution computed tomography (HRCT) images and retrieval of similar cases [6,7]. Detailed ground truth and a large number of cases allow to reliably evaluate and compare medical image processing algorithms for the defined tasks (i.e., benchmarks). Popular datasets such as Lena, Brodatz [8] or Iris[1] allow to qualitatively and quantitatively evaluate a large number of basic methods in image processing and machine learning and thereby established a *de facto* reference dataset.The popularity of these datasets is partly due to the fact that they reflect real-life challenges thus offering more credibility of the obtained results when compared to artificial datasets. It is also important that with such publicly available data sets results can be reproduced by other researchers and be compared to the state of the art, an important requirement in science. On the other hand, such databases also create opportunities for specialized studies and teaching. The cases with confirmed diagnoses constitute a knowledge base that can be used as diagnostic aid [9]. For instance, advanced browsing enabled by content-based image retrieval (CBIR) or multimodal case-based retrieval from large databases of cases with confirmed diagnoses can be highly valuable for radiologists with little experience in the domain [6,10,11]. In summary, a high-quality multimedia library of cases is valuable for:

- teaching,
- specialized descriptive studies,
- training and testing pattern recognition techniques,
- retrieving similar cases as diagnostic aid,

* Corresponding author at: University of Applied Sciences Western Switzerland (HES-SO), TechnoArk 3, CH-3960 Sierre, Switzerland. Tel.: +41 27 606 9033.
*E-mail address:* adrien.depeursinge@hevs.ch (A. Depeursinge).

[1] http://archive.ics.uci.edu/ml/datasets/Iris/, as of 24 May 2011.

0895-6111/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compmedimag.2011.07.003

- comparative performance analysis of medical image processing methods (i.e., benchmarks [12]).

However, the construction of a high-quality multimedia collection of cases is extremely time-consuming and expensive. It is often a bottleneck in studies on image-based CAD systems. The identification of the relevant cases, the consultation of the electronic health record (EHR) and picture archiving and communication systems (PACS) to gather the clinical parameters and the image series, the data entry as well as the database infrastructure and maintenance involve a large amount of work with a wide range of skills from medical knowledge to information technology (IT) expertise. To assess the high-quality of the data, several researchers and physicians have to be involved in the case selection process and the delineation of regions of interest (ROIs) to cope with the inter- and intra-observer variabilities, the latter being particularly important in radiology [13,14]. The agreement of the ethics committee has to be obtained before starting any investigations. The latter constitutes the required justification to access content of the EHR.

Depending on the studied diseases, cases are sometimes rare and are encountered casually even in large university hospitals. Incidentally, these diseases are the ones requiring reference databases to palliate the lack of experience due to their sparsity. Efforts from the European Union (EU) research programme came up with information technology (IT) infrastructures to cope with the difficulty of collecting rare cases with aneurysms in the AneurIST project[2] [15]. Multimedia data from six clinical centers within Europe was gathered. Anonymization of the patient data as well as images [16] is required as soon as the data leaves the medical institution [17].

A major observation when studying the state-of-the-art of texture-based CADs for lung tissue analysis in thin-section computed tomography (CT) is the lack of statistical significance of the measured performance as the CAD systems are most often evaluated on a small number of cases [18]. In this paper, the steps for building a multimedia database of cases with interstitial lung diseases (ILDs) and the current content of the created database are detailed. This database was built in the context of the Talisman[3] project [19].

### 1.1. Interstitial lung diseases

ILDs can be characterized by the gradual alteration of the lung parenchyma leading to breathing dysfunction. They regroup more than 150 histological diagnoses associated with disorders of the lung parenchyma [20]. The factors and mechanisms of the disease processes vary from one disease to another and the exact cause of many ILDs is still unknown [20]. Physical examination of a patient affected by ILD is frequently abnormal but with unspecific findings. The diagnosis of these pathologies is established based on the complete history of the patient, a physical examination, laboratory tests, pulmonary function testing (PFT) as well as visual findings on chest X-ray.

Images play an important role for confirming the diagnosis and patients may not require surgical lung biopsy when the clinical and radiographic impression is consistent with a safe diagnosis [21]. The first imaging examination used is the chest radiograph because of its low cost and weak radiation exposure. It also provides a quick overview of the entire chest. However, chest radiographs are normal in more than 10% of the patients with some forms of ILD and can provide a confident diagnosis in only 23% of the cases with lung diseases in general [22]. When the synthesis of this information arouses suspicions toward an ILD, HRCT imaging of the chest is often required to acquire a rapid and accurate visual assessment of the lung tissue. Indeed, the three-dimensional form of HRCT data avoids superposition of organs and provides an accurate assessment of the patterns and distribution of the lung tissue with a submillimetric resolution. It quickly became the gold standard imaging protocol for the diagnosis of diffuse pulmonary parenchymal diseases.

The most common histological diagnoses of ILDs according to [20,23] are detailed in Fig. 1. The associated lung tissue patterns in HRCT are listed in Table 2.

**Table 1**
The HRCT scanning protocol.

| | |
|---|---|
| Slice thickness | 1–2 mm |
| Spacing between slices | 10–15 mm |
| Scan time | 1–2 s |
| Lung shape | Inspirium |
| Contrast agent | None |
| Axial pixel matrix | 512 × 512 |
| *x*, *y* spacing | 0.4–1 mm |

#### *1.1.1. High-resolution computed tomography of the lung*

In 1972, the first commercial CT scanner created image series with relatively low resolution of an 80 × 80 pixel matrix in each axial slice. The numerical value of each pixel is related to the X-ray attenuation and is expressed in Hounsfield Units (HU). Modern scanners with multiple detectors use a helical scanning mode and can provide a 3D array of isotropic voxels with a submillimetric resolution. This protocol is called multidetector computed tomography (MDCT). The main drawback of this protocol is the high amount of radiation dose the patient is exposed to. To assess the visual appearance of healthy and pathological lung tissue, a submillimetric resolution is required but some areas can be skipped between the thin sections to limit the radiation exposure. This protocol is called HRCT and is the gold standard imaging protocol for diagnosing ILDs [24]. The technical specifications of the HRCT protocol are listed in Table 1.

HRCT is also more appropriate than magnetic resonance imaging (MRI) to assess the visual appearance of the lung tissue. Indeed, MRI is only sensitive to inflammatory changes of the pulmonary parenchyma as other tissue types have a low density of protons [25]. A comparison between HRCT and MRI is shown in Fig. 2.

#### *1.1.2. Lung tissue patterns associated with ILDs in HRCT*

The appearance and quantification of the types of lung tissue patterns in HRCT are very informative for establishing the differential diagnosis of an ILD. Table 2 lists 13 common histological diagnoses of ILDs, the associated HRCT findings as well as the region of the lungs where the disease is predominant. The visual aspects of the most common lung tissue patterns are depicted in Fig. 3. The taxonomy used to describe them often relates to texture properties. The term *fibrosis* is used in this work to describe all HRCT findings that are associated with the histological diagnosis "pulmonary fibrosis" and includes *reticulation*, *traction bronchiectasis*, *architecture distortion* and *honeycombing* [26]. As observed in Table 2, *ground glass* patterns are encountered in most of the ILDs and are thus non-specific. Therefore, the clinical context and other HRCT findings are required to orient the diagnosis.

---
[2] http://www.cistib.upf.edu/aneurist1/, as of 24 May 2011.
[3] TALISMAN: Texture Analysis of Lung ImageS for Medical diagnostic AssistaNce, http://www.sim.hcuge.ch/medgift/01_Talisman_EN.htm, as of 24 May 2011.