Interactive texture analysis in chest CT scans

Thessa T.J.P. Kockelkorn

Interactive texture analysis in chest CT scans
PhD thesis, Utrecht University, the Netherlands

The research described in this thesis was carried out at the Image Sciences Institute, University Medical Center Utrecht, Utrecht, the Netherlands, under the auspices of ImagO, the Graduate Programme Medical Imaging of the Utrecht Graduate School of Life Sciences.

ISBN/EAN 978-90-393-6706-3
Cover design and layout Thessa Kockelkorn
Printed by Uitgeverij BOXPress || Proefschriftmaken.nl

Interactive texture analysis in chest CT scans

Interactieve textuuranalyse in CT-scans van de thorax
(met een samenvatting in het Nederlands)

Proefschrift

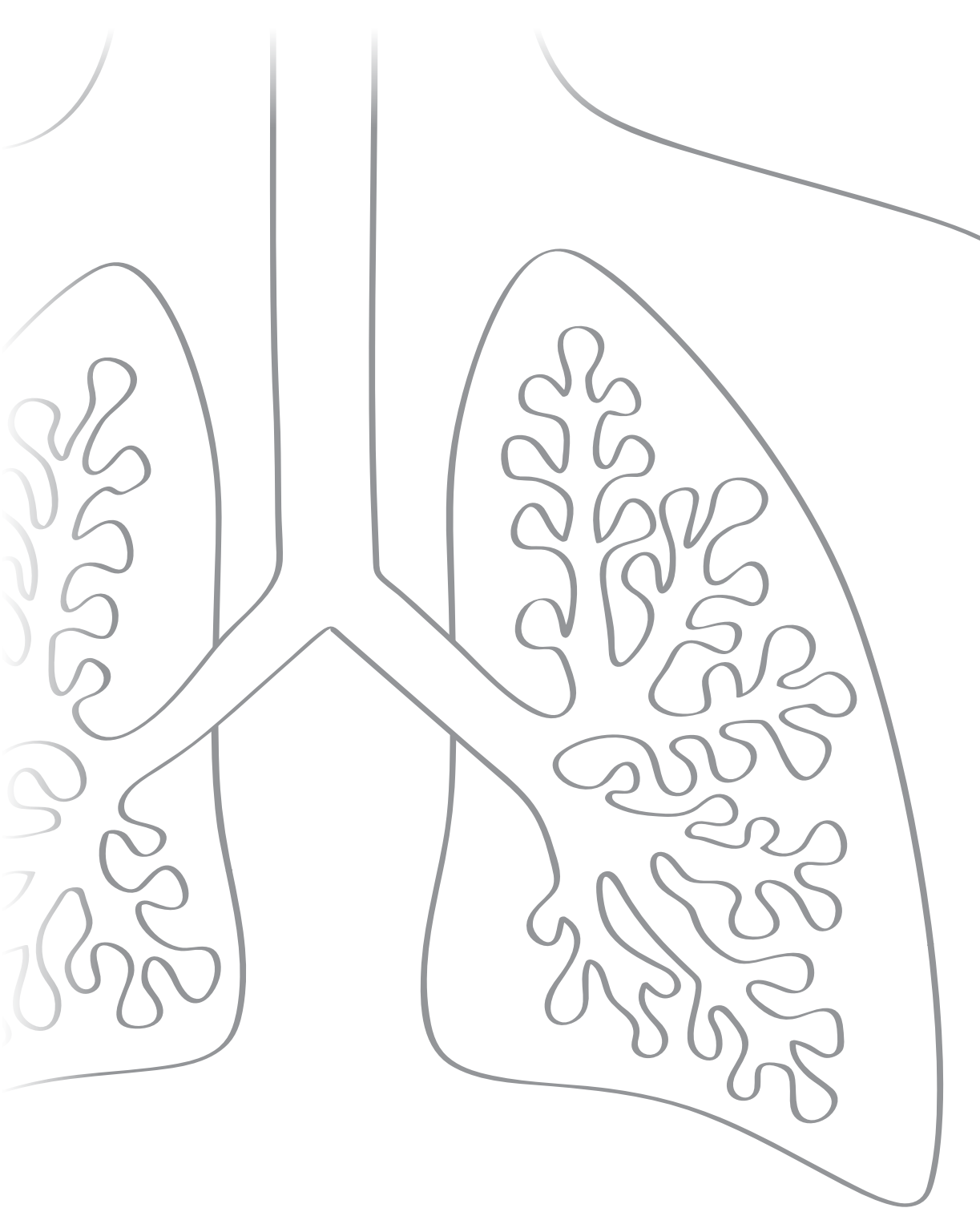
ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
donderdag 12 januari 2017 des middags te 2.30 uur

door

Thessa Theresia Johanna Petronella Kockelkorn
geboren op 15 december 1978 te Heerlen

Promotoren Prof. dr. ir. M.A. Viergever
Prof. dr. B. van Ginneken
Prof. dr. W.M. Prokop

Table of contents



Introduction

Medical imaging allows physicians to examine the inside of the living human body. For detailed inspection of the lungs, **computed tomography** is the technique of choice. This thesis discusses interactive **computer-aided diagnosis** methods that were developed to aid radiologists in their analysis of chest CT scans. The discussed methods use **classification** techniques to perform interactive texture analysis in CT scans of patients with **interstitial lung disease,** and to perform **interactive segmentation** of the lungs in chest CT scans.

In this chapter, the concepts highlighted in gray bold font are introduced, after which the remaining chapters are outlined.

Computed Tomography

Computed tomography (CT) is a medical imaging technique that can create detailed three-dimensional (3D) images of the inside of the live human body noninvasively. Since its introduction in the 1970s, CT has become one of the most commonly used medical imaging techniques. In 2013, 1.3 million CT scans were made in the Netherlands, which means numbers have more than tripled since the early 90s. Since 2001, the yearly rise was 9%. Twenty to 25% of these scans were thoracic CT scans (RIVM 2012; RIVM 2015).

Acquisition of CT scans involves an X-ray source that rotates around the subject, and detectors that measure the attenuation of the X-rays on the opposite side of the subject. These measurements are processed by a computer to construct slices: two-dimensional cross-sectional images of the subject. The final three-dimensional image is formed by stacking of the individual slices. Slices can be viewed individually. In addition, slices can also be extracted in other orientations (see Figure 1). Each voxel in a CT scan has a value that corresponds to the mean radiation attenuation of the tissue at that position. CT values of tissue, expressed in Hounsfield units (HU), are calculated relative to two fixed checkpoints: the CT value of water, which is set at 0 HU, and the CT value of air, which is set at -1000 HU. Lungs can be visualized well in CT scans. Air, with its low density, is depicted using dark shades, whereas the vascular tree and the walls of the larger bronchi appear as bright structures.

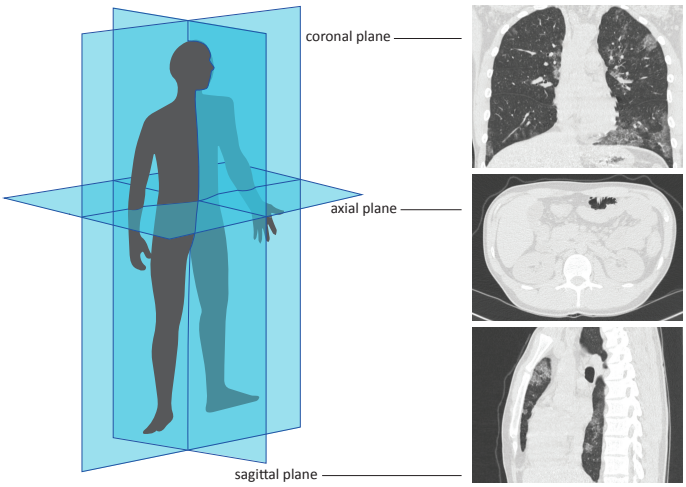


Figure 1. a. Schematical overview of axial, coronal, and the sagittal image planes. **b.** Examples of axial, coronal, and the sagittal chest CT slices.

Nowadays, CT scanning with sub-millimeter isotropic, i.e. equal in all three imaging planes, resolution is feasible. A thoracic CT scan with 1 mm slice spacing can be made within a single breath hold. For further information on the technical aspects of CT scanning, the reader is referred to the book by Kalender (2005).

Computer-aided diagnosis

The wide availability of medical imaging techniques comes at a price. Obvious side effects are increased radiation exposure, when ionizing radiation is used, and increased costs of health care. A less apparent consequence is the need for more radiologists for interpretation of the images. Alternatively, computer-aided diagnosis (CAD) can be used to assist physicians.

The field of CAD comprises technologies which are of help in the process of medical image interpretation. This help can be in the form of decreased time needed for a human observer to assess an image, or it can be in the form of increased accuracy in detection and quantification tasks (Van Ginneken, Schaefer-Prokop, and Prokop 2011). Areas in CAD research that have attracted much attention are mass detection in mammography, polyp detection in the colon, and lung nodule detection. An example of CAD applied to a quantification task is aortic and coronary calcium scoring in CT scans of the heart or chest.

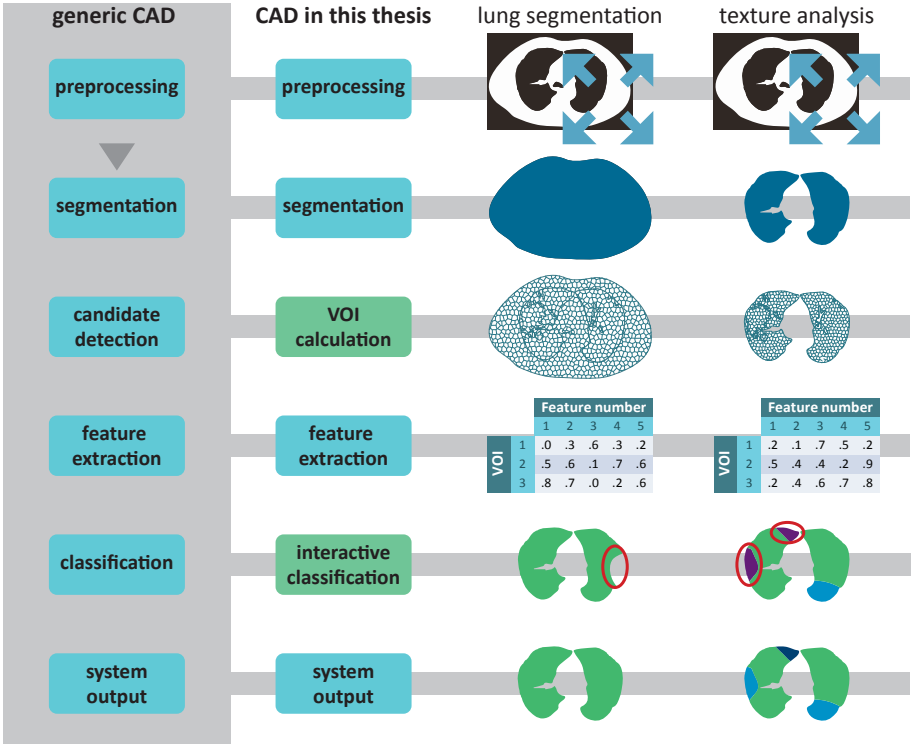


Figure 2. Components of a generic CAD system (left column), and components of the CAD systems described in this thesis (three columns on the right). VOI: volume of interest.

The review by Van Ginneken, Schaefer-Prokop, and Prokop (2005) outlines the components of a generic CAD system. The left column of Figure 2 is a graphic representation of these general components. The second column contains the translation of the generic components in this thesis. The third and fourth column exemplify each component for the two applications in this thesis: interactive texture analysis and interactive lung segmentation. Specifics for each component are discussed in the Methods sections of Chapters 2-4.

As a first step, scans are preprocessed, which in this thesis amounts to resizing of large scans. This is followed by automatic segmentation of the structure of interest, which are the chest in case of interactive lung segmentation, and the lungs in case of interactive texture analysis. In the generic CAD model, the third step is candidate detection. In both applications described in this thesis, this step is replaced by division of the structure of interest into volumes of interest (VOIs), as the aim of both applications is to assign a label to all voxels in the segmented structures. For all VOIs, features are calculated, which are used by a classifier in the next step to automatically label the VOIs. The classification approach in this thesis is interactive, which means that the classifier is trained by observers, while

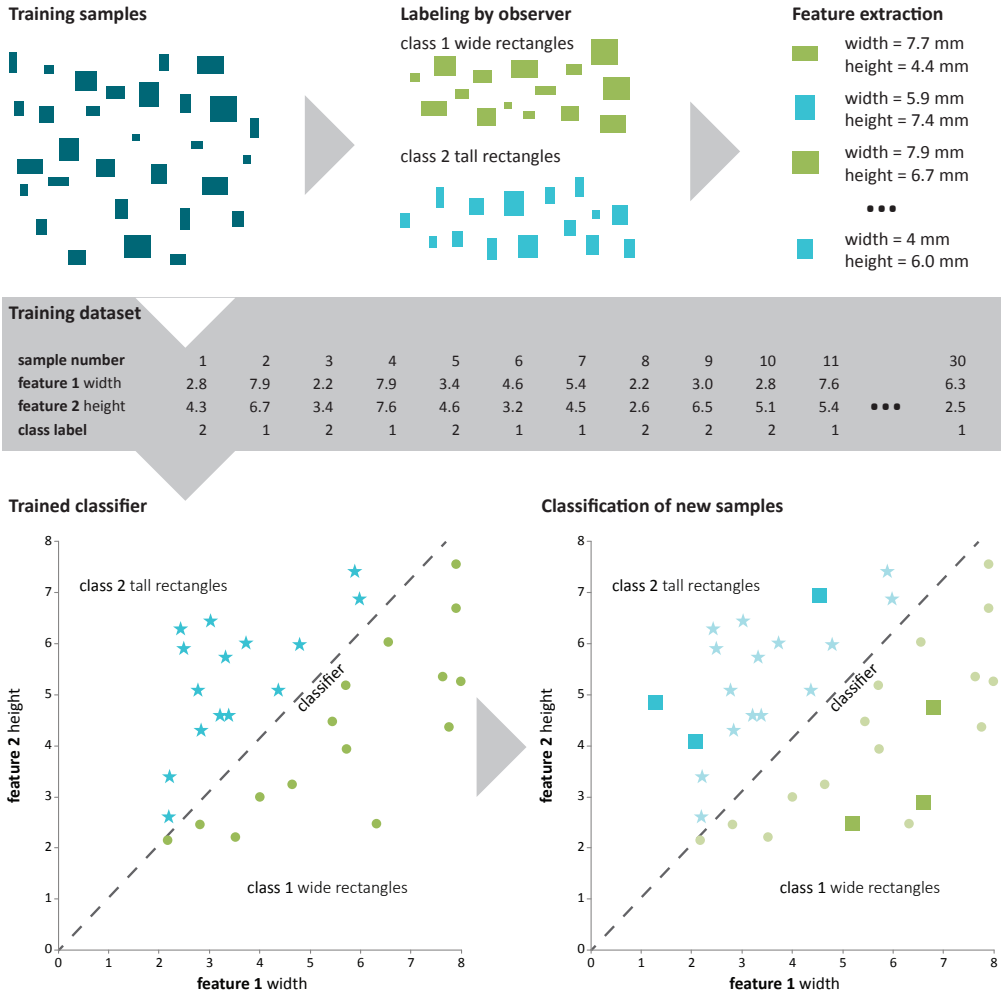


Figure 3. Example of automatic classification, in which a classifier is trained to distinguish tall and wide rectangles.

they perform an annotation or segmentation task. In the following paragraph, the concept of classification will be explained. Finally, the output is either a segmentation of the lungs or annotation of normal and abnormal textures in the lungs.

Classification

In order for a CAD system to deliver output, it needs to be trained. In the process of classification (see Figure 3), an algorithm, referred to as the classifier, is provided with examples of the elements that should be recognized. These examples, the training samples, need to contain two elements: features, which are numerical descriptions of the properties of the

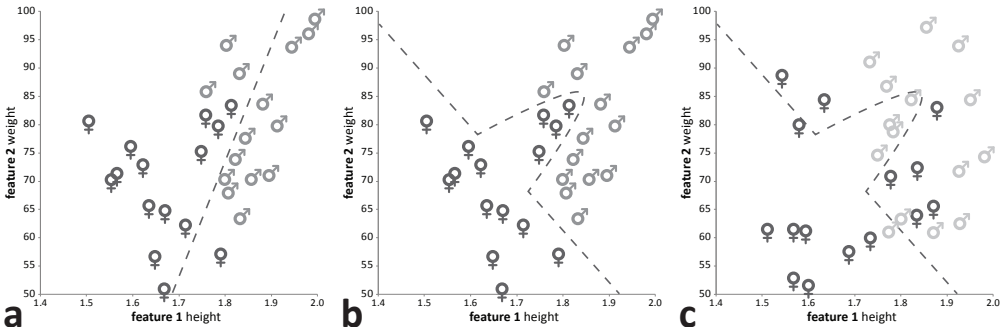


Figure 4. Scatter plots displaying the relation between height and weight in males and females. **a.** A linear classifier cannot separate the two classes. **b.** A non-linear classifier can separate the classes. **c.** As a result of overfitting, the non-linear classifier does not perform well on a test set of 30 new samples.

sample, and a label, which indicates the category to which the sample belongs. From the training samples, the classifier learns the relation between features and labels, which enables it to predict the labels of unseen samples. In Figure 3, the aim is to train a classifier to distinguish tall from wide rectangles. An observer assigns the labels 'tall' or 'wide' to a set of rectangles which serve as training data. Next, features need to be extracted, in this case the width and the height of the rectangles. The plot in the lower left-hand corner is a schematic representation of the classifier. The blue stars represent the tall rectangles, whereas the blue dots represent the wide ones. Both categories can be split by a linear classifier: all rectangles above the line $y=x$ are tall, and all rectangles below the line are wide. Once the classifier is trained, it can assign a label to rectangles which are not in the training dataset, based on their width and height.

Real-life classification tasks are more complicated than the toy example in Figure 3. Figure 4 is an example in which males and females are classified based on their height and weight. A linear classifier cannot separate both classes (Figure 4a), whereas a non-linear classifier can (Figure 4b). However, when this classifier is applied to a new dataset, as shown in Figure 4c, the results are disappointing. The trained classifier is not able to predict the sex of a subject reliably based on the subject's height and weight. Part of the problem is that the classifier in this case suffers from overfitting: instead of modeling the underlying relationship between height, weight, and sex, it modeled random noise.

In this case the two features that were chosen were not optimal. To solve this, more features could be added to better predict sex, such as hair length, favorite sport, and gravidity. By giving the classifier more information, it may be able to make a better distinction. But it is important to note that the key issue is not to collect more features, but rather to select the right features. In this particular example, the number of X chromosomes would suffice.

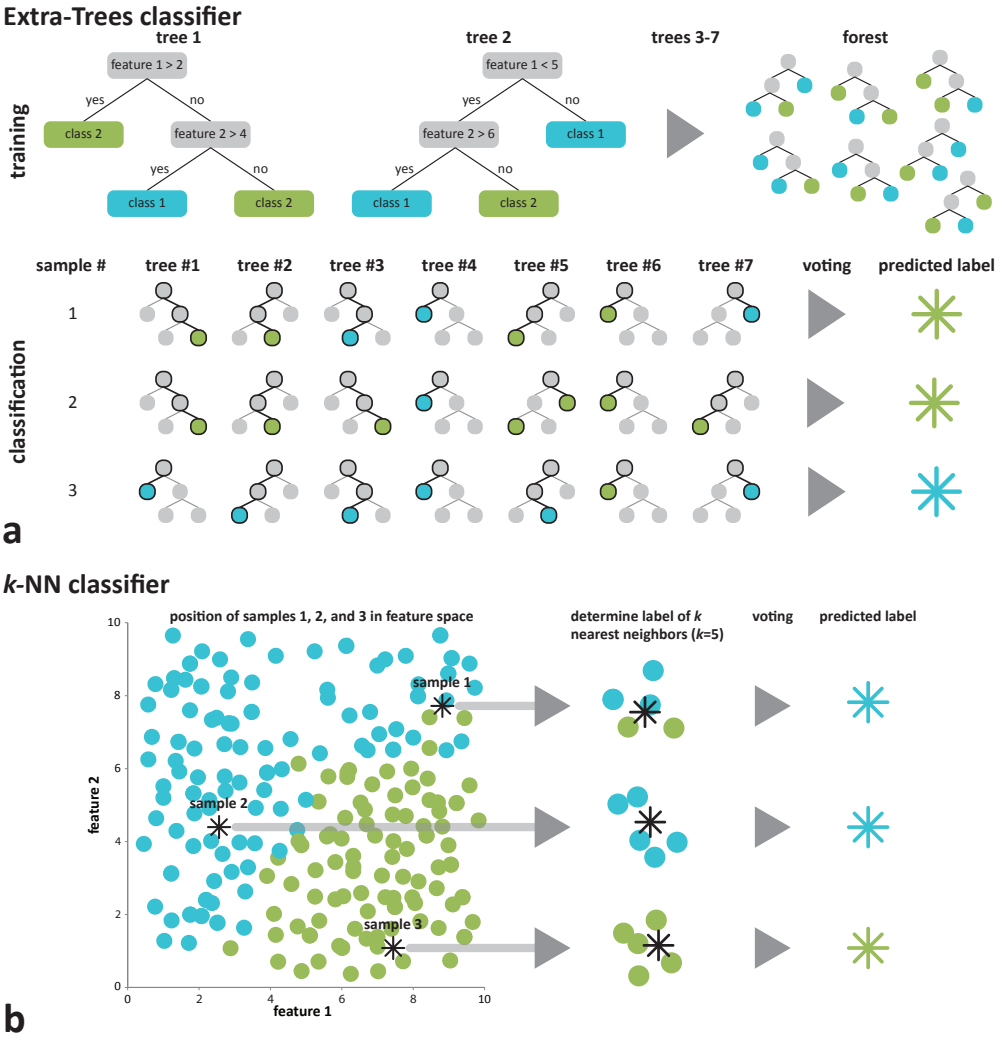


Figure 5. a. Schematic drawing of an Extra-Trees classfier. **b.** Schematic drawing of a *k*-NN classifier.

Two types of classifiers are used in the work described in this thesis: Extra-Trees classifiers (Geurts et al. 2006) and *k*-nearest neighbors (*k*-NN) classifiers (Duda, Hart, and Stork 2001). The Extra-Trees classifier is an example of a forest classifier. It consists of a user-defined number of decision trees. In the training phase, each tree is built using all training samples. Decision trees consist of nodes (the rounded rectangles in Figure 5a) and edges connecting them. At each node, the set of training samples is split. For this split, a random subset of the available features is selected. For each feature, a random threshold is chosen. Among these thresholds, the one which is best at dividing the group of training samples into the different classes is chosen as the splitting rule. Splitting of the nodes con-

tinues until each node contains only samples from one class. After training of the trees, new samples are classified by determining for each tree at which node they would end based on the chosen splitting rules. This node determines the label assigned to the sample. The final classification is determined by voting of all trees in the forest.

k-NN classifiers determine the label of a new sample by looking at the labels of the *k* nearest neighbors of the samples in feature space. For a classifier using two features, feature space can be depicted as a scatter plot (Figure 5b). The user defines the number of neighbors *k* that the classifier has to consider, most often an odd number to avoid ties in a two class problem. The new sample gets the label of the majority of its neighbors.

Choice of training data

A key issue in each training process, for humans as well as for computers, is the selection of appropriate training material. The labels are assigned to the training samples by human observers. By making these annotations, they determine the reference standard from which the classifier learns - if samples are incorrectly labeled, the classifier may learn incorrect relationships between features and labels. Obtaining reliable ground truth data is therefore crucial, but not trivial, when building a CAD algorithm.

First, the scope of the problem, and hence the training dataset, should be determined. A larger scope requires more training data to capture the variation in the population of interest. A smaller scope requires less training samples, but the resulting classifier may not be trained to handle samples outside the scope. Second, decisions should be made on the size of the training dataset. For complex problems, in general more training is necessary to obtain good classification results. As labeling is usually a labor-intensive task, the benefits of having a large training dataset should be weighed against the costs of obtaining it.

Above issues are universal, but additional challenges arise if the ground truth is difficult to establish. This may happen if there is little consensus between different observers on the correct labels of samples. A classifier trained on contradicting information may not be trained effectively. In addition, its performance cannot be measured accurately, since it is unclear to which standard its predictions should be compared.

Interstitial lung disease

The lungs are organs that perform gas exchange: they transfer oxygen from the air into the blood stream and release carbon dioxide from the blood stream into the air. Figure 6 shows a schematic drawing of the lungs (left). Inhaled air flows through the trachea, via

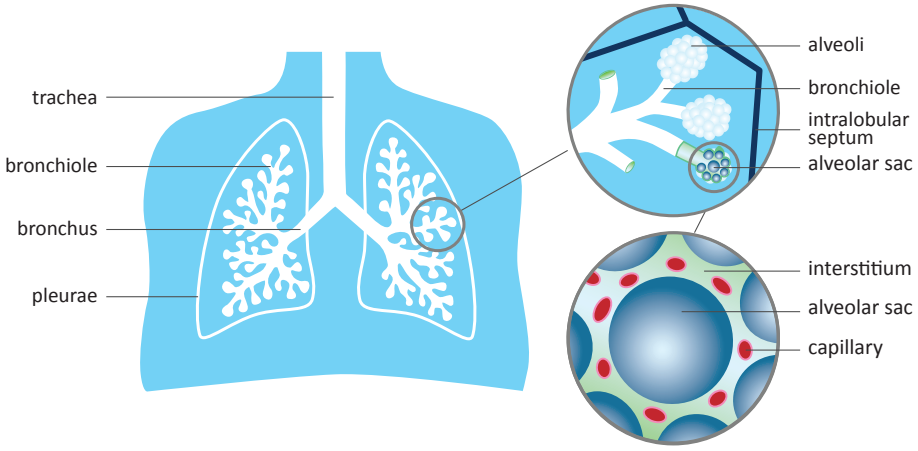


Figure 6. Schematic overview of the lungs, bronchioles, and alveoli.

the bronchi, into the bronchioles, to arrive in the alveoli (top right), where gas exchange takes place. The pulmonary interstitium is the network of tissue supporting the alveoli (bottom left). The lungs are surrounded by two pleurae, with a thin layer of pleural fluid between them to allow the pleurae to slide against each other during respiration.

Interstitial lung disease (ILD) is the collective noun for a diverse group of over 200 diseases that affect the pulmonary interstitium. Patients usually present with breathlessness and reduced exercise tolerance. For around 65% of all patients, the cause of their condition is unknown. Examples of these diseases are sarcoidosis, interstitial pulmonary fibrosis (IPF), non-specific interstitial pneumonia (NSIP), and ILDs secondary to connective tissue disease or collagen-vascular disease. For the other 35% of patients, the etiology is known. Examples of these ILDs are asbestosis, silicosis, extrinsic allergic alveolitis, iatrogenic ILD caused by drugs or radiation, and post-infectious ILD (Gibson et al. 2013).

Prognosis and preferred treatment vary widely, depending on the exact type of ILD. Therefore, making the correct diagnosis is of pivotal importance. Diagnosis requires an interdisciplinary approach, using a combination of pulmonary function tests, imaging, and laboratory tests (Wells et al. 2008). CT scans are known to be an important factor in making ILD diagnoses (Aziz et al. 2006). When lung tissue is injured, it reacts in a predictable way. Therefore, various disease processes result in similar changes in the lung tissue, which in turn manifest themselves similarly in CT scans. When making the diagnosis, the distribution of imaging characteristics within the lungs should be taken into account (Jawad et al. 2012).

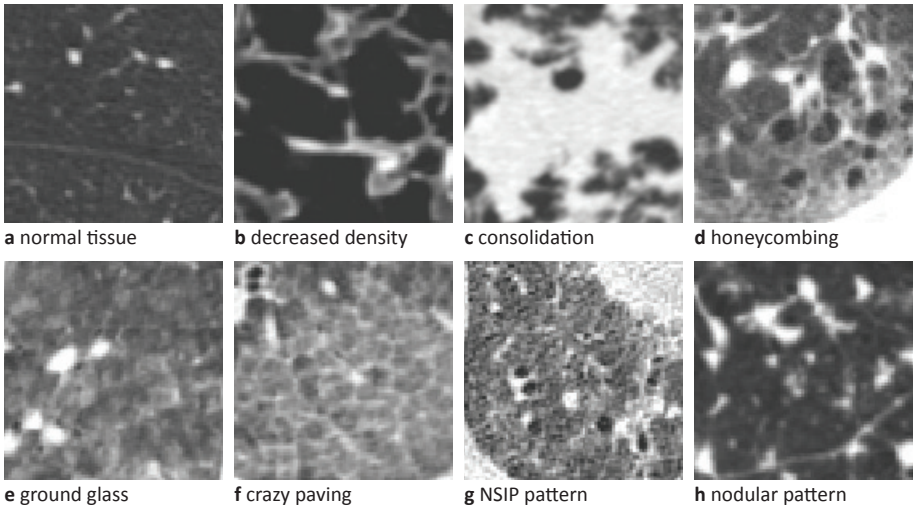


Figure 7. Examples of the different lung textures that were used in lung tissue analysis in this thesis.

In this thesis, 8 types of normal and abnormal lung tissue are distinguished in chest CT scans of ILD patients. These tissue types are referred to as textures. Figure 7 shows examples of each texture. Normal tissue (a) is tissue without any abnormalities. The decreased density class (b) encompasses tissue with decreased attenuation values as compared with normal lung parenchyma. These areas of low attenuation can be, but are not necessarily surrounded by walls. The consolidation class (c) consist of areas with increased attenuation values. The underlying bronchovascular structures are no longer visible. Often, air-filled airways are visible as dark tubular structures amidst fluid-filled alveoli. Honeycombing (d) is made up of cysts with varying diameters (0.3-1.0 cm), stacked in layers. These cysts have relatively thick walls and lower attenuation values as compared with normal lung tissue. The ground glass class (e) displays increased attenuation values, with visible bronchovascular structures. Various disease processes may present themselves as ground glass in CT scans and in most cases, these processes are reversible. Crazy paving (f) is characterized by linear pattern, which is superimposed on ground glass opacity. It resembles a pattern of paving stones with irregular shapes. The NSIP pattern (g) also contains ground glass, but with architectural distortion of lung parenchyma, traction bronchiectasis or irregular lines. Traction bronchiectasis is irreversible abnormal dilatation of the bronchial tree. Finally, nodular pattern (h) is characterized by sharply defined areas of increased attenuation, of 1-4 mm in diameter, in a random or paralymphatic distribution. Nodules can also have branching structures, referred to as tree-in-bud (Jawad et al. 2012; Kockelkorn et al. 2016).

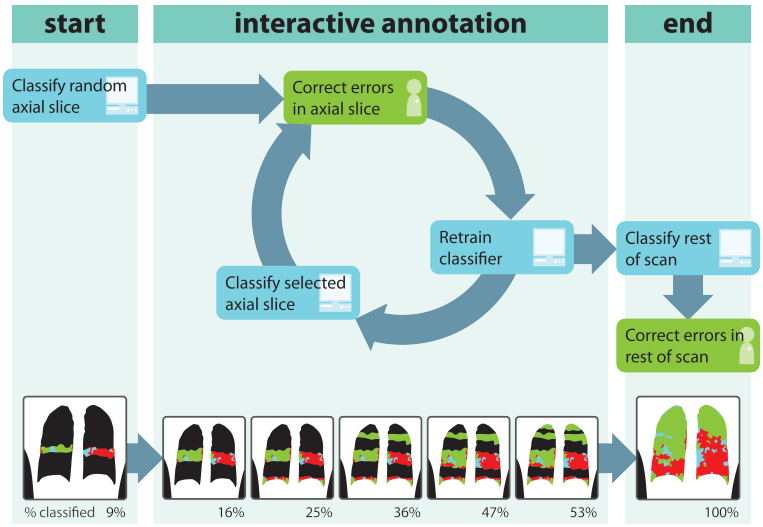


Figure 8. Flowchart of interactive annotation. The core of the process is the annotation cycle, in which an observer corrects automatic classification results of VOIs in an axial slice, after which these VOIs and their corresponding labels are added to the training data, which is then used for retraining the classifier. In this way, the classifier learns to distinguish the different textures present in the scan by means of the corrections and approvals of the observer. After a predefined part of the scan has been annotated in this manner, the classifier is retrained and all unlabeled VOIs are classified and inspected by the observer.

Automatic classification of lung textures in CT scans of ILD patients is an active field of study. Various 2D automatic methods have been proposed (e.g. Uppaluri et al. 1999, Park et al. 2009, Depeursinge et al. 2010, Depeursinge et al. 2012, Huber et al. 2011, Huber et al. 2012, Vasconcelos et al. 2015). Typically, ILD classification methods are trained and tested on hand-drawn regions of interest (ROIs), or automatically generated square ROIs. Previous work in our group indicated that automatically generated ROIs, containing one type of texture, better resembled hand-drawn ROIs than square regions (Sluimer et al. 2006). In the work described in this thesis, this approach was extended to 3D and embedded in an interactive annotation framework.

Figure 8 is a flowchart of the interactive annotation framework, in which a classifier is iteratively trained to classify pre-calculated VOIs. After initialization, a human observer inspects the labeled VOIs intersecting with an axial slice and relabels incorrectly classified VOIs. A classifier is then trained using the VOIs in the slice inspected by the observer as training data. Classification results for the VOIs intersecting with a second axial slice are shown to the observer for correction, after which the VOIs in the second slice are added to the training data and the classifier is retrained. This cyclic process of correction, retraining, and classification continues until a predefined percentage of the lung tissue has been reviewed, or until the observer considers the classifier optimally trained. The classifier is

trained once more, after which the remaining VOIs are classified. Finally, the observer inspects the entire scan to correct any remaining errors.

Alternatively, observers can opt to label all VOIs manually, or to correct automatic classification results of all VOIs at once, without iteratively training a classifier.

Interactive segmentation methods

With their low density compared to the surrounding tissue, lungs are easily visible on CT. Many algorithms for automatic lung segmentation have been developed (see van Rikxoort and van Ginneken 2013 for a review). For diseased lungs, automatic segmentation is not always possible for diseased lungs. Lung tissue containing pathology may have higher density values, thus resembling the surrounding tissue. On the other hand, some dense abnormalities that are actually situated in the lung pleurae can push lung tissue aside. Since these abnormalities are located in areas where the lungs normally are, they can be mistaken for lung pathology. Examples of such abnormalities are pleural effusion, in which fluid accumulates between the pleurae, and pleural tumors. In these cases, automatic lung segmentation methods are likely to fail and an interactive approach is called for.

Interactive segmentation methods utilize automatic segmentation techniques to decrease the amount of user interaction as compared with manual segmentation. At the same time, user involvement ensures that high segmentation accuracy is possible, even for difficult cases. As the term 'interactive' spans the entire spectrum between completely manual and completely automatic, there are many possible ways to implement interactive segmentation, with various degrees of user interaction. The main forms of interaction are initialization of the methods, checking of the automatic results, and correction of the results if necessary. The exact type of interaction depends on the automatic segmentation method that underlies the interactive approach. Zhao and Xie (2013) provided an overview of the different interactive segmentation methods used for medical images.

In this thesis, interactive segmentation is performed in a manner similar to interactive classification in Figure 8. The main differences are that the chest instead of the lungs are divided into VOIs, and that only two types of textures are distinguished: lung tissue and non-lung tissue.

Thesis outline

When proven fully automatic methods are not available, interactive annotation methods can be useful alternatives for manual annotation. The aim of the work described in this thesis is the development of interactive tools to facilitate annotation of chest CT scans. Three possible applications are considered: annotation of normal and abnormal lung textures in CT scans of ILD patients, lung segmentation in abnormal human and animal CT scans, and interactive lung segmentation and subsequent aeration analysis in chest CT scans of patients admitted to the intensive care unit (ICU).

In **Chapter 2**, the interactive annotation framework in Figure 8 is applied to 10 scans of ILD patients. These scans are part of the ILD database, an initiative of the St. Antonius Hospital in Nieuwegein and the University Medical Center in Utrecht. In this database, CT scans of patients with ILD are collected, with their corresponding diagnoses. The scans in this database are not annotated and can be used for research purposes. One automatic and two interactive annotation protocols are evaluated using software that simulated observer behavior. The interactive methods differ in the way interactive annotation is initiated: by a classifier trained on data from previously annotated scans or by a heuristic approach. The protocols are compared in terms of the percentages of VOIs that need to be relabeled to obtain completely annotated scans.

Chapter 3 discusses extensions of the interactive annotation framework introduced in Chapter 2. Also in this chapter, scans from the ILD database are used. Several approaches for using training data of previously annotated scans are evaluated, as well as methods by which observers can transfer their knowledge on the textures present in the scan to the annotation environment, other than by relabeling VOIs. Results are evaluated in terms of the percentages of VOIs that are correctly classified. In addition, different methods for selection of the axial slices that are shown to the observer are compared to determine which method requires the smallest amount of user interaction.

Chapter 4 describes interactive lung segmentation. As the application of the interactive framework is not limited to human scans, not only abnormal human lungs, but also lungs of pigs and mice are segmented. The resulting lung segmentations are evaluated in terms of required user interaction, and in terms of segmentation accuracy.

For many patients admitted to the ICU, mechanical ventilation is a life-saving intervention. In lungs that are already injured, mechanical ventilation may cause further damage if too much air is administered (Gattinoni and Pesenti 2005). A common method to estimate the required ventilation settings is based on the patient's height. The main problem with

this method is that it is not able to adapt to the degree to which the lungs are still able to perform gas exchange. Thoracic CT scans enable assessment of lung injury and can provide additional information for determining ventilation parameters (Gattinoni et al. 2006, Caironi et al. 2010). In **Chapter 5**, interactive lung segmentation and subsequent lung texture analysis in CT scans of patients admitted to the ICU are used to make an estimate of the total lung capacity and the lung volume that is able to perform gas exchange. Results are compared to estimation of total lung volume based on patient height.

Finally, **Chapter 6** summarizes the main results of this thesis and discusses the possible further embedding of interactive annotation tools into CAD systems.

References

Aziz ZA, Wells AU, Bateman ED, Copley SJ, Desai SR, Grutters JC, Milne DG, Phillips GD, Smallwood D, Wiggins J, Wilsher ML, Hansell DM. Interstitial lung disease: effects of thin-section CT on clinical decision making. *Radiology.* **2006**;238(2):725-33.

Caironi P, Cressoni M, Chiumello D, Ranieri M, Quintel M, Russo SG, Cornejo R, Bugedo G, Carlesso E, Russo R, Caspani L, Gattinoni L. Lung opening and closing during ventilation of acute respiratory distress syndrome. *Am J Respir Crit Care Med.* **2010**;181(6):578-86.

Depeursinge A, Iavindrasana J, Hidki A, Cohen G, Geissbuhler A, Platon A, Poletti P-A, Müller H. Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization. *Journal of digital imaging* **2010**;23(1): 18–30.

Depeursinge A, Van de Ville D, Platon A, Geissbuhler A, Poletti P-A, Müller H. Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames. *IEEE Transactions on Information Technology in Biomedicine* **2012;**16(4): 665-675

Duda RO, Hart PE. and Stork, D.G. *Pattern classification* (New York: John Wiley and Sons) 2nd ed., **2001**

Gattinoni L, Pesenti A. The concept of "baby lung". *Intensive Care Med.* **2005**;31(6):776-84.

Gattinoni L, Caironi P, Valenza F, Carlesso E. The role of CT-scan studies for the diagnosis and therapy of acute respiratory distress syndrome. *Clin Chest Med.* **2006**;27(4):559-70.

Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn* **2006**;63:3-42.

Huber MB, Bunte K, Nagarajan MB, Biehl M, Ray LA, Wismüller A. Texture feature ranking with relevance learning to classify interstitial lung disease patterns. *Artificial Intelligence in Medicine.* **2012; 56**(2):91–97.

Huber MB, Nagarajan MB, Leinsinger G, Eibel R, Ray LA, Wismüller A. Performance of topological texture features to classify fibrotic interstitial lung disease patterns. *Medical Physics* **2011;38**(4): 2035–2044.

Gibson GJ, Loddenkemper R, Sibille Y, Lundbäck B. Interstitial lung diseases. In: *The European Lung White Book* (Sheffield: European Respiratory Society) **2013**.

van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology* **2011**;261(3):719-32.

Jawad H, Chung JH, Lynch DA, Newell JD Jr. Radiological approach to interstitial lung disease: a guide for the nonradiologist. *Clin Chest Med.* **2012**;33(1):11-26. doi: 10.1016/j.ccm.2012.01.002.

Kalender WA. *Computed Tomography: Fundamentals, System Technology, Image Quality, Applications.* (Erlangen, Germany: Publicis Corporate Publishing) 2nd ed., **2005**

Kockelkorn TTJP, de Jong PA, Schaefer-Prokop CM, Wittenberg R, Tiehuis AM, Gietema HA, Grutters JC, Viergever MA, van Ginneken B. Semi-automatic classification of textures in thoracic CT scans. *Phys Med Biol.* **2016**;61(16):5906-24.

Park SO, Seo JB, Kim N, Park SH, Lee YK, Park BW, Sung YS, Lee Y, Lee J, Kang SH. Feasibility of automated quantification of regional disease patterns depicted on high-resolution computed tomography in patients with various diffuse lung diseases. *Korean J Radiol.* **2009**;10(5):455-63.

Rijksinstituut voor Volksgezondheid en Milieu (RIVM), Medische Stralingstoepassingen **2012**.

Rijksinstituut voor Volksgezondheid en Milieu (RIVM) [internet]. Trends in het aantal CT-onderzoeken. **2015** Available from: http://www.rivm.nl/Onderwerpen/M/Medische_Stralingstoepassingen/Trends_en_stand_van_zaken/Diagnostiek/Computer_Tomografie/Trends_in_het_aantal_CT_onderzoeken. Cited August 21, 2016.

van Rikxoort EM, van Ginneken B. Automated segmentation of pulmonary structures in thoracic computed tomography scans: a review. *Phys Med Biol.* **2013**;58(17):R187-220.

Schwartz M, King Jr TE. *Interstitial Lung disease.* (Shelton, CT: People's Medical Clearing House) 5th ed., **2011**.

Sluimer IC, Prokop M, Hartmann I, van Ginneken B. Automated classification of hyperlucency, fibrosis, ground glass, solid, and focal lesions in high-resolution CT of the lung. *Med Phys.* **2006**;33(7):2610-20.

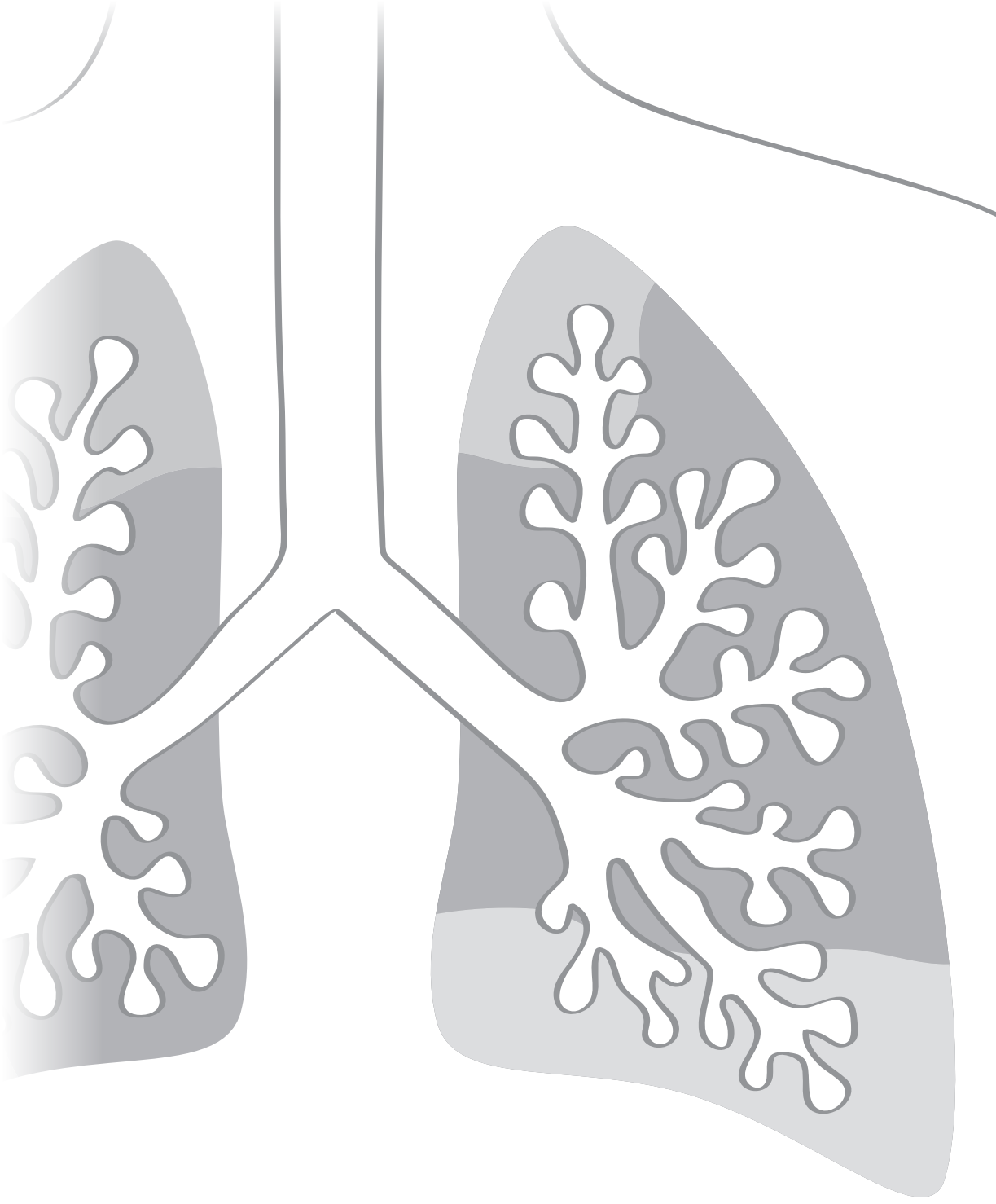
Ulzheimer S, Flohr T. Multislice CT: Current Technology and Future Developments. In: *Multislice CT* (Berlin Heidelberg: Springer) 3rd ed., **2009**;3-23

Uppaluri R, Hoffman EA, Sonka M, Hunninghake GW, McLennan G. Interstitial lung disease: A quantitative study using the adaptive multiple feature method. *Am J Respir Crit Care Med.* **1999**;159(2):519-25.

Vasconcelos V, Barroso J, Marques L, Silva JS. Enhanced Classification of Interstitial Lung Disease Patterns in HRCT Images Using Differential Lacunarity. *Biomed Res Int.* **2015**;2015:672520.

Wells AU, Hirani N, on behalf of the British Thoracic Society Interstitial Lung Disease Guideline Group, a subgroup of the British Thoracic Society Standards of Care Committee, in collaboration with the Thoracic Society of Australia and New Zealand and the Irish Thoracic Society. Interstitial lung disease guideline: the British Thoracic Society in collaboration with the Thoracic Society of Australia and New Zealand and the Irish Thoracic Society. *Thorax.* **2008**;63 Suppl 5:v1-58.

Zhao F, Xie X. Overview on interactive medical segmentation. *Annals of the BMVA.* **2013**:2013(7);1–22.



Semi-automatic classification of textures in thoracic CT scans

Thessa TJP Kockelkorn
Pim A de Jong
Cornelia M Schaefer-Prokop
Rianne Wittenberg
Audrey Tiehuis
Hester Gietema
Jan C Grutters
Max A Viergever
Bram van Ginneken

Based on: Physics in Medicine and Biology, 081915 (2016)

Abstract

The textural patterns in the lung parenchyma, as visible on computed tomography (CT) scans, are essential to make a correct diagnosis in interstitial lung disease. We developed one automatic and two interactive protocols for classification of normal and seven types of abnormal lung textures. Lungs were segmented and subdivided into volumes of interest (VOIs) with homogeneous texture using a clustering approach. In the automatic protocol, VOIs were classified automatically by an Extra-Trees classifier that was trained using annotations of VOIs from other CT scans. In the interactive protocols, an observer iteratively trained an Extra-Trees classifier to distinguish the different textures, by correcting the mistakes that the classifier makes in a slice- by-slice manner. The difference between the two interactive methods was whether or not training data from previously annotated scans was used in classification of the first slice. The protocols were compared in terms of the percentages of VOIs that observers needed to relabel. Validation experiments were carried out using software that simulated observer behavior. In the automatic classification protocol, observers needed to relabel on average 58% of the VOIs. During interactive annotation without the use of previous training data, the average percentage of relabeled VOIs decreased from 64% for the first slice to 13% for the second half of the scan. Overall, 21% of the VOIs were relabeled. When previous training data was available, the average overall percentage of VOIs requiring relabeling was 20%, decreasing from 56% in the first slice to 13% in the second half of the scan.

1 Introduction

Interstitial lung disease (ILD) is a diversity of approximately 200 rare diseases that affect the lung tissue around the pulmonary alveoli. With so many different ILDs, each with its own preferred management, making the correct diagnosis is crucial, but also complicated. Diagnosis is usually based on a combination of pulmonary function testing, imaging and lung tissue analysis, with a pivotal role for chest CT scans (Aziz et al. 2006).

ILDs may present themselves in different ways in chest CT scans. Common signs are hyperlucency, fibrosis, honeycombing and ground glass. The nature of the abnormalities present in a scan, their quantity, and their distribution need to be assessed in order to determine the type and severity of the ILD the patient is suffering from. CT scans can also be beneficial in later stages of disease, to monitor disease progression or treatment response. Two factors complicate the interpretation of CT scans in ILD patients. First, the imaging features of different ILDs are partly overlapping. Second, many of the individual

ILDs are rare, and therefore not often seen by radiologists, especially in smaller hospitals. Moreover, even for experienced radiologists, chest CT analysis in ILD patients is a labor-intensive and complex task in large 3D volumes. Automatic and interactive methods to analyze textures in thoracic scans may provide means to make this task less complicated and less time-consuming.

1.1 Related work

A number of 2D automatic classification methods for pathological textures in the lungs of ILD patients have been proposed previously (Uppaluri et al. 1999, Park et al. 2009, Depeursinge et al. 2010, Depeursinge et al. 2012, Huber et al. 2011, Huber et al. 2012). Much of this work has focused on finding optimal features and classifiers for the problem at hand. With the increase of volumetric CT scans being made, comparable systems for analysis of 3D volumes of interest (VOIs) have also been proposed. A major advantage of these systems is that they allow quantification of abnormalities. Xu and coworkers described a system for classification of emphysema, ground-glass, honeycombing, normal smokers' parenchyma and normal non-smokers' parenchyma. Sensitivity ranged from 73% to 93% and specificity from 90% to 99%, depending on the classified tissue and the classifier used (Xu et al. 2006). Zavaletta et al. have used a similar 3D approach for classifying VOIs into one of the following classes: honeycombing, reticular, ground glass normal and emphysema. They also extended their analysis to whole lungs, discriminating normal tissue, reticular tissue and honeycombing (Zavaletta et al. 2007).

While these previously described approaches indicate that texture analysis in thoracic CT scans is possible, two major drawbacks should be noted. First, the definition of the VOIs is not trivial. VOIs can be drawn manually by an expert, but this is rather labor-intensive. Alternatively, automatically generated cubic regions can be used. Their main disadvantage is that they will not follow the boundaries between two different types of texture precisely. As a consequence, either some VOIs will contain more than one type of texture or VOIs containing multiple textures should be excluded from further processing. Second, suitable training data for these systems may not always be available. Manual annotation requires a substantial amount of time. In addition, using annotations from other observers and other institutes is not an optimal solution. The first reason for this is that in clinical scans, acquisition parameters and the use of exogenous contrast may vary. Using training samples from scans which were acquired with different parameters may negatively influence classification results. The second reason is that low interobserver agreement, defined as the percentage of ROIs or VOIs to which two observers independently assign the same label, is a known challenge in ILD texture analysis. Uppaluri et al. used 31×31 pixel ROIs in lung

parenchyma of patients with different lung diseases. Experienced observers had to label these ROIs as one of six patterns. In two readings, observers were blinded to the diagnosis of the subject. Interobserver agreement was 49% for the first reading and 52% for the second reading. In the third reading, observers were provided with the patient's diagnosis. Interobserver agreement was 54% (Uppaluri et al. 1999). Sluimer described experiments in which two observers assigned one of six labels to lung ROIs containing homogeneous texture. One observer annotated these ROIs once, the other observer twice. Interobserver agreement was 77%. Intra-observer agreement, defined as the percentage of ROIs receiving the same label from one observer at two readings, was 89% (Sluimer et al. 2006). In one of our previous studies, two observers annotated 3D lung VOIs with homogeneous texture using six labels. They agreed on 51% of the labels (Kockelkorn et al. 2010).

1.2 Contributions

The goal of our project is to develop a system that facilitates fast annotation of normal and abnormal textures in thoracic CT scans. Given that individual observers tend to annotate in different ways, and scanning parameters vary, we did not aim for universally accepted classification results. Instead, we aimed to develop a system that would adapt to the annotation preferences of the observer who uses it and to the specifics of the scan under consideration. Two aspects make our approach novel. First, we used irregularly shaped VOIs with homogeneous texture. Previous work in our group showed that in 2D, irregularly shaped regions of interest (ROIs) showed a better resemblance to hand- drawn outlines than square regions (Sluimer et al. 2006). We extended this approach to 3D volumetric scans (Kockelkorn et al. 2010). Annotation is performed by assigning a texture label to each VOI in the lungs. In this way, the disadvantages of manually drawing outlines of textures on the one hand and of using VOIs that contain more than one type of texture on the other hand are overcome. Second, the annotation environment we developed uses an interactively trained classifier. This means that the classifier learns to distinguish the different textures present in a scan from the observer. Annotation takes place in a slice-by-slice manner. At the beginning of the annotation process, the system is still untrained and proposes classifications that are relatively inaccurate. The observer corrects the mistakes, and after each slice, the classifier is retrained using this input, thus becoming increasingly capable of recognizing the different textures.

We chose to implement software that simulates interactive annotation sessions, instead of having radiologists performing the experiments. In this way, we were able to test different classification strategies and repeat experiments if necessary, without having to ask observers to invest large amounts of time.

Table 1. Patient and scan characteristics.

Scan number	Patient age	Patient sex	Number of VOIs	In-plane resolution (mm)	Slice spacing (mm)	Peak voltage (kV)	Tube current (mA)
1	67	female	2021	0.605	0.8	120	120
2	20	female	2786	0.574	0.8	120	217
3	33	male	3084	0.688	0.8	120	144
4	45	male	1647	0.873	0.8	120	144
5	42	female	2040	0.658	0.8	120	144
6	61	male	1234	0.781	1.0	120	192
7	73	male	1947	0.729	0.8	120	144
8	46	female	1432	0.621	0.8	120	90
9	74	female	2092	0.586	0.8	120	90
10	24	male	2818	0.645	0.7	120	206

2 Materials

Ten volumetric thoracic CT scans from ILD patients were collected retrospectively from the St. Antonius Hospital (Nieuwegein, the Netherlands). Scans were selected to contain a variety of texture patterns, which are described below. The diagnoses of the patients in these scans were usual interstitial pneumonia (4 cases), non- specific interstitial pneumonia (NSIP, 2 cases), sarcoidosis, hypersensitivity pneumonitis, pulmonary Langerhans cell histiocytosis, and pulmonary alveolar proteinosis (1 case each). All scans were acquired on a Philips Mx8000 IDT scanner (Philips Healthcare, Eindhoven, the Netherlands). Patient and scanning protocol parameters are summarized in table 1. Images were acquired at maximal inspiration with patients in supine position and reconstructed to 512×512 matrices. In-plane resolution ranged between 0.57×0.57 and 0.87×0.87 mm with slice spacings between 0.7 and 1.0 mm. The peak tube voltage was 120 kV for all cases; tube current ranged between 90 and 217 mA.

The following definitions were used for the textures (Figure 1):

Normal tissue Lung tissue without any abnormalities

Decreased density Decreased density compared with normal lung parenchyma, with or without surrounding walls

Consolidation Increased lung density, in which underlying structures are no longer visible

Honeycombing Cystic destruction of subpleural lung parenchyma: there are cysts of varying diameter (0.3-1.0 cm) in several layers and cysts share relatively thick walls

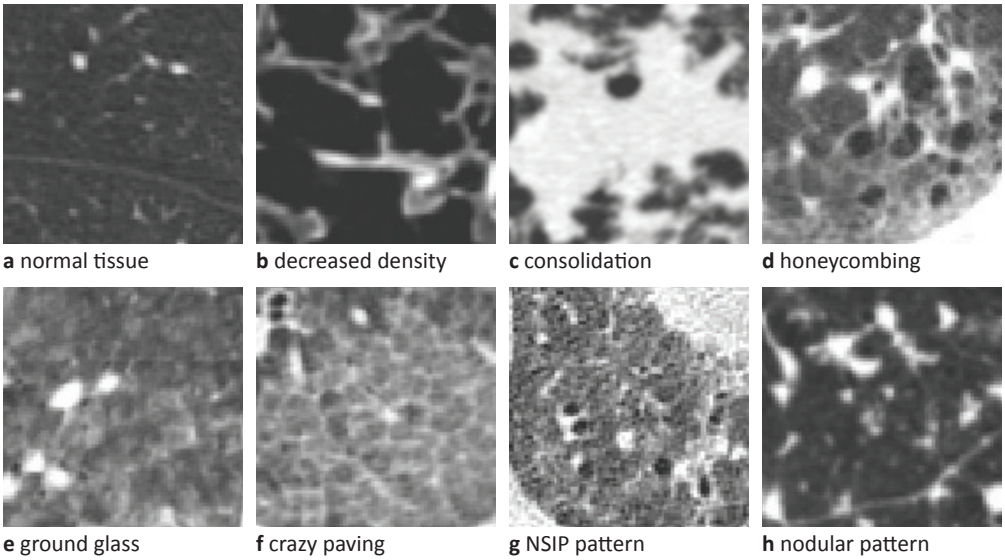


Figure 1. Examples of the different lung textures.

Ground glass Increased lung density, in which underlying structures are still visible

Crazy paving Regular pattern of ground glass with reticular pattern

NSIP pattern Ground glass with architectural distortion, traction bronchiectasis or irregular lines

Nodular pattern Sharply defined nodular densities (1-4 mm) in a random or paralymphatic (paraseptal) distribution. Nodules can also have branching structures (tree-in-bud)

3 Methods

The (semi-)automatic classification methods described in this paper all have the same starting point. First, the lungs in the CT scans under consideration are automatically divided into VOIs (paragraph 3.1). Then, features are calculated to represent each VOI numerically for classification purposes. Details on the features, classifier and training are given in paragraph 3.2. We describe one automatic and two interactive protocols for annotation of the VOIs (paragraph 3.3). These different protocols are tested in the experiments listed in paragraph 3.4.

3.1 VOI creation

In all scans, the lungs were automatically segmented (Hu et al. 2001, van Rikxoort et al. 2009). Subsequently, the lungs were subdivided into roughly spherical VOIs containing

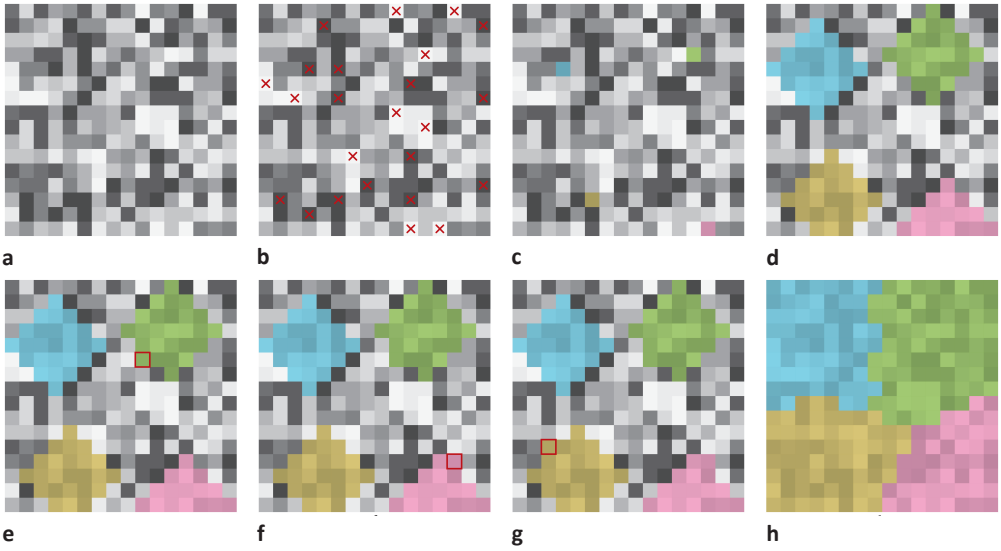


Figure 2. Schematic overview of the VOI creation process

voxels with similar density values using a clustering approach (Kockelkorn et al. 2010, Kockelkorn et al. 2014). The process of VOI creation is depicted schematically in Figure 2. In short, scans are downsampled to 256×256 matrices with isotropic voxels and blurred using a Gaussian kernel (σ = 1 voxel). Local minima and maxima in the lungs, with a minimal distance of 8 voxels, are selected as seeds for the VOIs. Around these seeds, initial VOIs are formed by adding all lung voxels within a radius of ⅓× the minimum seed distance. VOIs are further grown in a competitive way. For each voxel neighboring a VOI, a score is calculated. This score *D* indicates the dissimilarity between the VOI and the neighboring voxel as follows:

$$D = \left|(H_v - \overline{H})\right| + C \times d^2 \qquad (1)$$

where *Hv* is the density of the neighboring voxel in Hounsfield units (HU), $\overline{H}$ is the average density value in the VOI, and *d* denotes the distance in mm from the voxel to the center of the neighboring VOI. *C* is the relative weight given to the squared difference in distance. In this study, a *C* of 3 was used. This number was obtained by visual inspection: the resulting VOIs are in general roughly spherical, while they contain a limited range of density values. The neighbor with the lowest score is added to its neighboring VOI and for all new neighboring voxels, dissimilarity scores are calculated. The process of adding voxels to VOIs continues until all lung voxels belong to a VOI, after which the scan was resized to its original dimensions. On average, lungs were divided into 2210 VOIs (range: 1234-3084).

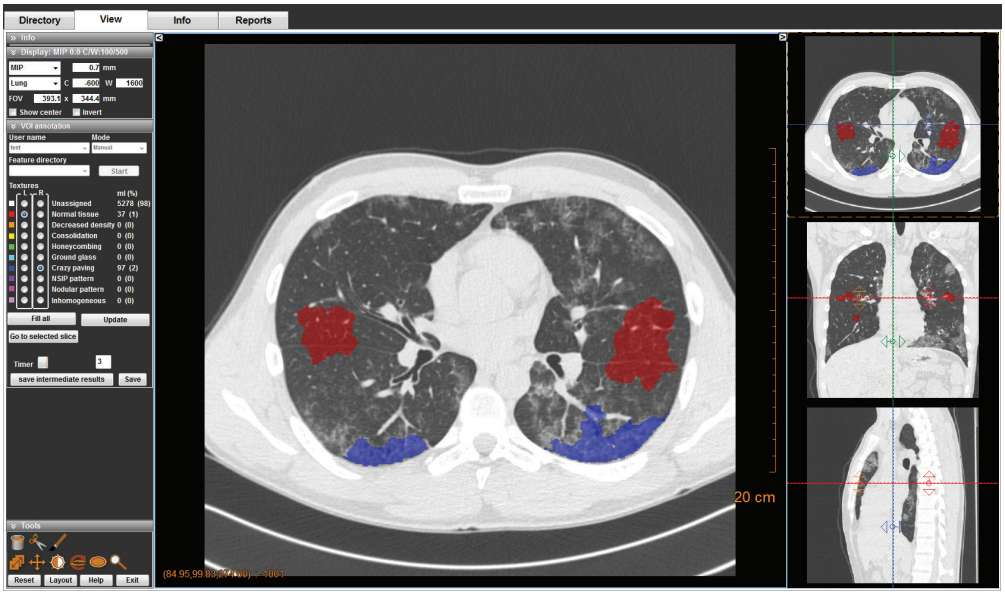


Figure 3. Screenshot of the annotation environment in interactive mode. The axial, coronal and sagittal images on the right hand side show the 3D orientation of the slice in the main window.

3.2 Features, classifier and training

Once the lungs are divided into VOIs, these VOIs can be labeled. This labeling can be done manually, by asking an observer to assign a texture category to each VOI, or automatically. For automatic labeling, VOIs need to be represented by a set of features.

For each VOI, 72 features were calculated. Scans were filtered using a Gaussian, a Laplacian, a gradient magnitude, and three Hessian eigenanalysis-based filters. These six filters were applied at three scales (σ = 1, 2 and 4 voxels), which resulted in 18 filtered images per scan. In each of these filtered imaged, we calculated average, standard deviation, skew and kurtosis of the densities per VOI. This resulted in 72 numbers that were used as features. Rotationally invariant features were chosen, since the lung textures that were classified do not have specific orientations. We used the Extra-Trees algorithm (Geurts et al. 2006) as a classifier in all experiments. This supervised ensemble method builds a number of randomized decision trees. In this work, 99 trees were used. Each tree is built using all available training samples, in contrast to other tree-based ensemble methods, which use a bootstrap replica. For each node in a tree, a random subset of features is used. In this work we used 10 features per node. For each of these features, a threshold is drawn at random. The best of these thresholds is selected as the splitting rule for the corresponding node. The minimum number of samples after a split was set to 1. All parameter choices were derived from the values described by Geurts et al. (Geurts et al. 2006).

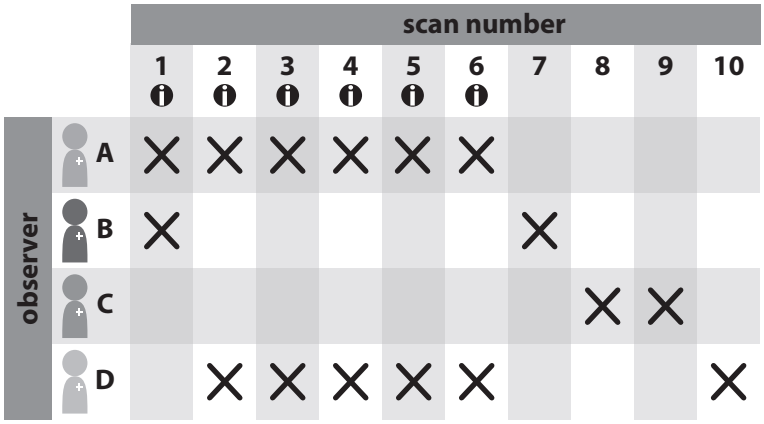


Figure 4. Table indicating which scans have been annotated by which observer. A cross in cell (*i, j*) indicates that the VOIs in scan *i* have been labeled by observer *j*. Scans that were annotated by two observers independently, thus yielding two annotated datasets, are indicated by i.

For training the Extra-Trees classifier, two different training schemes were used in this study. The first scheme uses training data from other, previously annotated scans. A leave-one-scan-out setup was used, which means that for classification of each annotated dataset, training data was obtained from the other nine scans. From all annotated VOIs in these nine scans, 500 VOIs per texture category were randomly selected as training data. If a texture category contained less than 500 training samples, all training samples were used. For decreased density, the number of selected training samples was lower than 500 for the classification of two annotated datasets. In those cases, 319 samples were used. The number of training samples available for consolidation was on average 338 (range: 173-386). The classifier trained on these samples is referred to as $c_{previous}$. The other training scheme uses training data from the scan under consideration. In the beginning, when no VOIs are labeled, the training dataset is empty. During interactive annotation, all VOIs that have received a label that is approved by the observer are added to the training dataset. The Extra-Trees classifier is then retrained. This Extra-Trees classifier is referred to as $c_{current}$.

3.3 Annotation

The software for our experiments was developed by our group and written in C++. A screenshot of the annotation software is displayed in Figure 3. Annotations can be made in manual and in interactive mode. To each mouse button, a texture can be assigned using the radio button menu on the left hand side. The user can assign labels to single VOIs by clicking or to multiple VOIs by dragging in the axial slices of the CT scan shown in the center panel. Labels are shown as a colored semi-transparent overlay on top of the CT scan.

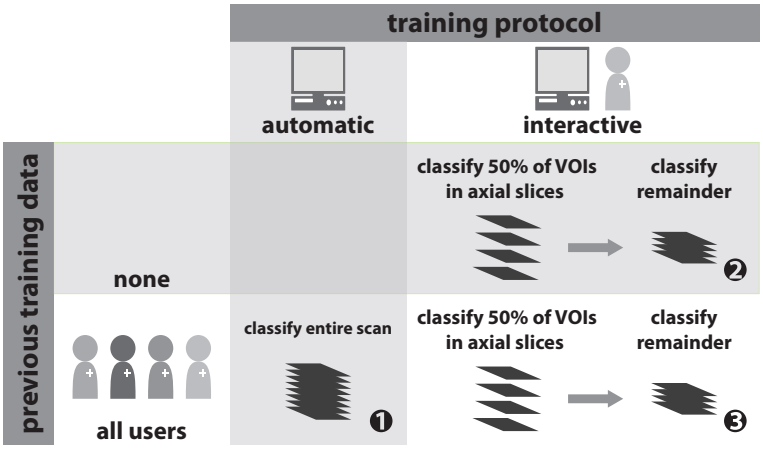


Figure 5. Schematic overview of the three classification protocols tested in this paper. The completely automatic (protocol 1, left hand side) is described in paragraph 3.3.2. Here, training data from other scans annotated by all observers is used. In the other two protocols, (on the right hand side, described in paragraph 3.3.3), interactive classification is used. In protocol 2 (top), the classifier is trained using data from the scan that is being annotated by the observer. Protocol 3 uses training data from other scans for classification of the first slice. The remaining slices are classified using training data from the current scan.

3.3.1 Manual labeling of VOIs

In manual mode, the observer has to assign a label to each VOI. In these experiments, four radiology residents, with 2 to 8 years of experience with chest CT reading, were asked to annotate a subset of the ten selected scans as indicated in Figure 4. Four scans were annotated by one observer, and six scans were independently annotated by two observers. The VOIs in the scans that were annotated twice were the same for both observers. This yielded sixteen annotated datasets. Scans annotated by two observers were treated as two different ground truths. For each of these ground truths, separate automatic and interactive classification experiments were performed. The radiologists had to distinguish the following textures: normal tissue, decreased density, consolidation, honeycombing, ground glass, crazy paving, NSIP pattern and nodular pattern. If a VOI contained more than one type of texture, observers were asked to label it as heterogeneous. These VOIs were not included in further analysis. On average, manual annotation of all VOIs took 62 minutes per scan.

3.3.2 Automatic labeling of VOIs

The three classification protocols described in this work are graphically represented in Figure 5. Automatic classification (protocol 1, left hand side in Figure 5) was performed in a leave-one-scan-out set-up. For annotation of each of the sixteen annotated datasets (ten

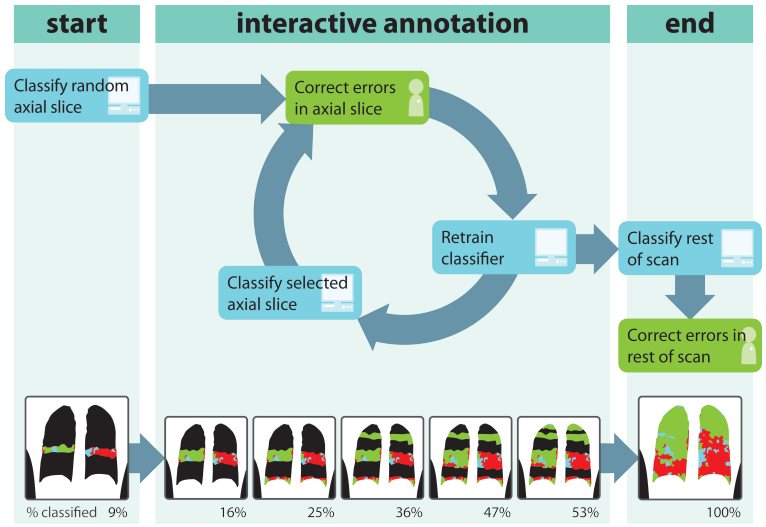


Figure 6. Schematic overview of the interactive VOI-based annotation procedure. Green rounded rectangles represent user actions, blue rounded rectangles represent computer actions. The images at the bottom of the figure depict the annotations at different time points in the annotation procedure. The left image represents annotations after classification of the first axial slice; the right image represents the final annotated scan. In the smaller images in the middle, results after each classification and correction cycle are drawn. The numbers underneath the images are the percentages of VOIs in the scan that have been annotated at each time point.

scans, of which six were annotated twice), the Extra-Trees classifier $c_{previous}$ was trained using the annotations of the nine remaining scans. No user interaction was required.

3.3.3 Interactive labeling of VOIs

The interactive annotation procedure (protocols 2 and 3 in Figure 5) is schematically depicted in Figure 6. Steps executed by the human observer are indicated by green rounded rectangles, steps executed by the computer are indicated by blue rounded rectangles. First, the interactive software randomly selects an axial slice to be annotated. If previous training data is available, $c_{previous}$ is used for classification of the first slice. In case no previous training data is used, all VOIs in the first slice are assumed to contain normal tissue, since this is the texture with the highest prior probability.

The labels assigned to the VOIs in the selected slice are displayed as a semi- transparent overlay. The core of the interactive annotation procedure is a cyclic process. Each cycle consists of the observer correcting classification errors in the selected slice, followed by retraining of the classifier $c_{current}$ on all annotated VOIs in the scan so far, and subsequently the selection and classification of a new axial slice. At any point in the interactive annotation process, observers can scroll through the CT scan. Thus, they can view all slices that

the newly labeled VOIs span when reviewing an axial slice and take all voxels of the VOIs into consideration when deciding on the label. Observers can also scroll through unseen slices and through slices they have reviewed already. In the latter case, they can change the labels of VOIs that were assigned in previous cycles. If the label of a VOI is adapted, this is also reflected in the training data for the Extra-Trees classifier: when retraining the classifier, the updated label is used instead of the one that was previously assigned. This cyclic process continues until at least 50% of the VOIs has received a label that is approved by the user. If annotation of 50% is reached in a slice, all VOIs in that slice are classified and corrected. Therefore, at the end of interactive training, over 50% of the VOIs have received a label. The remaining VOIs are classified, and all labels are shown. The observer now corrects the last classification errors. After correction, all lung voxels have received their definitive label.

From active learning, it is known that the amount of information gained by asking the observer to either label unannotated samples or to correct sample labels depends on the selection of samples that are presented to the observer (Settles 2009). We made the following choices in order to optimize this selection and minimize the amount of user effort for annotation of a scan. First, we chose to ask the user to correct classification results of all VOIs intersecting with an axial slice before retraining the classifier. This reduces the total time the user has to wait for classification results as compared with an approach in which the user corrects classification results of individual VOIs, in which case the classifier has to be retrained after the review of each VOI. In addition, radiologists are used to interpreting CT scans in a slice-based manner rather than inspecting individual VOIs. Second, disease processes in ILD can be localized. Therefore, it is important to sample all parts of the lungs when training the interactive classifier. We divided the part of the CT scan containing the lungs into 5 areas in axial direction, each containing an equal number of slices. These parts are numbered 1 through 5 from the apex to the base of the lungs. In the first cycle, a slice is chosen at random from the entire scan. In the second cycle, a slice from the 3rd area was chosen, then from the 5th area, the 2nd, the 4th, the 1st, again from the 3rd area, and so on. This resulted in sampling of different parts of the lungs, as can be seen in Figure 6. Third, to determine which slice was annotated within one area, we calculated the cumulative uncertainty of the classifier over all n unassigned VOIs in a slice:

$$U = \sum_{v=1}^{n} (1 - c_v) \qquad (2)$$

U is the uncertainty score, c_v denotes the confidence of the classifier for VOI v, and 1-c_v is the uncertainty. In each cycle, the slice with the largest uncertainty score was chosen, assuming that this slice contained the most valuable information for the classifier.

We simulated two protocols for interactive annotation. Simulated interactive labeling was essentially the same as the process depicted in Figure 6. All automatic steps, in the blue rounded rectangles, are executed in the same way as in the regular interactive protocols. The two steps executed by human observers, in the green rounded rectangles, are changed. Instead of asking a human observer to review the classification results, classification results are compared to the manual annotations of the scan under consideration. If the label of a VOI predicted by the classifier is the same as the manually assigned label, the classification is correct. If the automatically and the manually assigned label were different, the classification is incorrect. In this case, the VOI is relabeled. In both cases, the classified and reviewed VOI is added to the training data of $c_{current}$.

3.4 Experiments

Three experiments, schematically depicted in Figure 5, were performed. Classification of each dataset, as described below, was simulated 10 times, to average the effect of random selection of the training samples (in protocols 1 and 3) and the effect of randomly selecting the first slice for classification (in protocols 2 and 3).

3.4.1 Automatic classification

In the automatic classification protocol (protocol 1), experiments were conducted in a leave-one-scan-out set-up. One annotated dataset was used as test data, while training data for each texture category was randomly selected from all other scans, annotated by all observers.

3.4.2 Interactive classification without previous training data

In protocol 2, scans were classified interactively without the use of previous training data for classification of the first randomly chosen slice. Since no training data was available for this slice, we used an heuristic approach and all VOIs were assumed to contain normal tissue. The following slices were classified using the labeled VOIs in the slices already reviewed by the observer as training data. At least 50% of all VOIs in the dataset were classified in this slice-by-slice manner. The remaining VOIs were classified all at once, using the classifier that was trained on the first half of all VOIs.

3.4.3 Interactive classification with previous training data

In protocol 3, annotated datasets were also classified interactively. For classification of the first randomly chosen slice, training samples were randomly selected from all other scans using annotations from all observers. The experiments in protocol 3 were therefore also

conducted in a leave-one-scan-out set-up. The slices after the first slice were classified using the classifier trained on the labeled VOIs in all previously reviewed slices. After at least 50% of all VOIs had been reviewed in the slice-by-slice manner, the remaining VOIs were classified simultaneously, using a classifier trained on the first 50% of the VOIs in the dataset.

3.5 Evaluation

For all protocols, results were assessed by comparing classification results of all VOIs in a scan to manual annotations. For simulated interactive annotation, results were recorded for each classified axial slice, for the remaining VOIs after annotation of at least 50% of all VOIs and for the entire scan.

The primary outcome measure for this research was the percentage of VOIs that needed relabeling after classification by the Extra-Trees classifier. To this end, we calculated classifier accuracy as follows:

$$accuracy = \frac{n_{\text{correct}}}{n_{\text{total}}} \times 100\% \qquad (3)$$

where $n_{correct}$ is the number of correctly classified VOIs and n_{total} is the total number of classified VOIs. Inter-observer agreement was calculated by dividing the number of VOIs which received the same label from two observers by the total number of VOIs annotated by both observers. The percentage of VOIs requiring relabeling was calculated as 100% – accuracy. Accuracy was calculated for each classified slice in each dataset, for classification of the rest of the scan after at least 50% of the VOIs had been annotated interactively, and for all VOIs. Results were calculated per dataset. To examine the effect of adding more training data from the scan under consideration on classification accuracy, average accuracy over all datasets was also calculated for using 0% (i.e. the first slice), 0 – 10%, 10 – 20%, 20 – 30%, 30 – 40%, 40 – 50% and > 50% of the VOIs in a scan as training data.

In addition, classification results were expressed in terms of sensitivity and specificity to each texture category. Sensitivity was defined as:

$$sensitivity = \frac{TP}{P} \times 100\% \qquad (4)$$

where *TP* stands for the number of true positives and *P* is the total number of positives per texture category.

Specificity was calculated as

$$specificity = \frac{TN}{N} \times 100\% \qquad (5)$$

where *TN* denotes the number of true negatives and *N* the total number of negatives for each texture category. Sensitivity and specificity were calculated over all datasets.

4 Results

4.1 Example of manual, automatic and interactive annotation

In this study, six of the ten scans were annotated by two observers independently. On average, interobserver agreement was 63%. Figure 7 shows an example slice (a) that was annotated by Observer A (panel c) and Observer D (panel d). This slice was chosen since it displays low inter-observer agreement. If previous training data is used for classification of this slice (panel b), the classifier predicts the presence of tissue from all 8 texture categories. This slice was also classified using the interactive annotation approach. Since both observers annotated the slice in a different way, the resulting interactive classifiers also show different behavior. Classification results after training on at least 50% of the VOIs in the scan are shown for Observer A (panel e) and for Observer D (panel f). These last two results indicate that the interactive annotation set up adapts well to the specific annotation behavior of the observer who trained it.

4.2 Classification accuracy

Figure 8 displays classifier accuracy for interactive classification without (a) and with (b) previous training data. Accuracy is shown as a function of the percentage VOIs in the scan used as training data. Each bubble displays the accuracy of classification of one slice in a dataset, averaged over the ten times each experiment was performed. The area of the bubble is proportional to the average number of VOIs that were classified in the slice. Classification accuracy, averaged over all datasets and all repetitions of the experiments, is displayed by the blue bars. The thin left bar indicates classifier accuracy of the first slice, when no VOIs in the scan were used as training data. The second bar displays the average accuracy if 0-10% of the VOIs were used as training data, the third bar shows the accuracy for 10-20% training data and so on. The last bar displays the accuracy when at least 50% of the VOIs in a scan were used as training data, which corresponds to classification of the rest of the scan after complete interactive training. For the interactive annotation approach in which no previous training data was used, 36% of the VOIs in the first slice were correctly labeled as normal tissue. When previous training data was used for classification of the first slice, average accuracy over all annotated datasets was 44%. For classification of all following slices, training data from already labeled VOIs in the scan was used. For both interactive protocols, average classification accuracy increased with an increasing

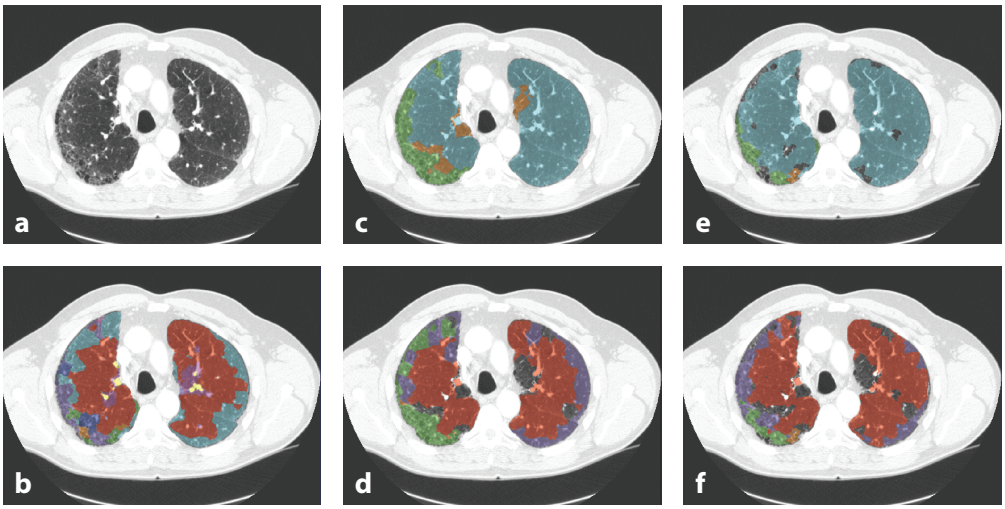


Figure 7. Example of an axial slice (a) with (b) automatic classification results (c) annotations by Observer A (d) annotations by Observer D (e) interactive classification results after training on at least 50% of all VOIs in the scan annotated by Observer A and (f) interactive classification results after training on at least 50% of all VOIs in the scan annotated by Observer D.

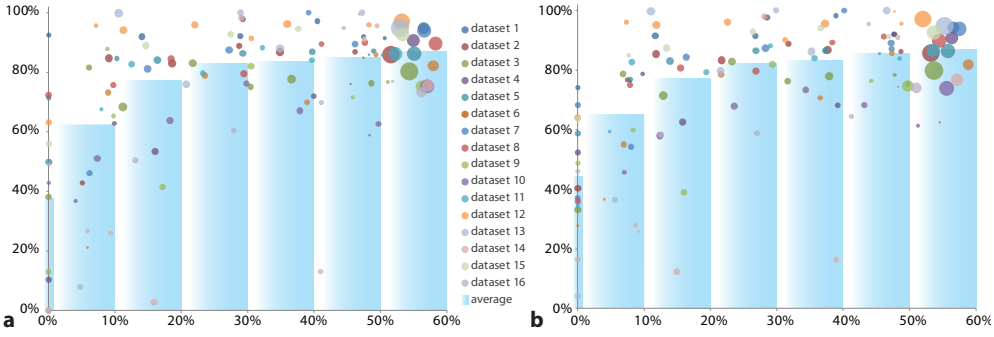


Figure 8. Bubble plot charts displaying classifier accuracy per annotated dataset. Left panel: results for interactive annotation without previous training data; right panel: results for interactive annotation with previous training data. Each data point displays the percentage correctly classified VOIs in an axial slice, as a function of the percentage of VOIs in a scan used as training data. Results were averaged over the 10 repetitions of interactive classification for each scan. The area of the bubbles is proportional to the number of VOIs that were classified. The bars indicate classifier accuracy averaged over all annotated datasets. The thin left bar is the average accuracy for classification of the first slice, where 0% of the VOIs in the dataset were used as training data. The second bar displays the average accuracy when 0-10% of the VOIs in the scan were used as training data, the third 10%-20% and so on. The last bar is the average accuracy for classification of the rest of the scan after training on at least 50% of all VOIs.

Table 2. Percentages of classified VOIs of which labels needed to be corrected for the different annotation protocols. In columns, from left to right, are results for classification of the first slice, classification of the remaining VOIs after interactive annotation of at least 50% of all VOIs and for classification all VOIs in a scan. All percentages for the interactive protocols are averaged over the ten times each experiment was performed and over all sixteen annotated datasets. Standard deviations are given in parentheses.

Protocol	First slice	After training on 50% of VOIs	Overall
Automatic (1)	-	-	58% (19%)
Interactive (2)	64% (33%)	13% (8%)	21% (11%)
Interactive with training data (3)	56% (21%)	13% (8%)	20% (10%)

number of training samples. Classification of the remaining VOIs after training on 50% of the VOIs in a scan was done with 87% accuracy for the interactive protocols, both without and with previous training data. The bubbles show that classification accuracy varies per slice and per dataset.

Table 2 summarizes the results of the different classification strategies in terms of user effort required for complete annotation of the lung tissue in a thoracic CT scan. If the scan is automatically classified, the user needs to correct on average 58% of the automatically assigned labels. After interactive training on at least 50% of all VOIs in a scan, the user needs to correct the label of 13% of the remaining VOIs in the scan. When interactively annotating a complete scan in the absence of previous training data, the user needs to correct the label of only 21% of all VOIs on average. If previous training data was used for classification of the first slice, the user needs to correct on average 20% of the labels of all VOIs.

In the six scans that were annotated by two observers, we were able to divide the VOIs into two groups: VOIs that received the same label from both observers (easy VOIs) and VOIs that received different labels (hard VOIs). For both groups, we compared the percentages of VOIs requiring relabeling in the entire scan, in protocols 1, 2 and 3. Table 3 lists the results for classification of all VOIs. In the six scans considered in this analysis, the overall percentages of VOIs that require relabeling are similar to the percentages obtained for classification of all sixteen annotated datasets. For all three protocols, the percentages of easy VOIs requiring relabeling are lower than the percentages of hard VOIs requiring relabeling. In both interactive protocols, the percentage of hard VOIs that is incorrectly classified is twice the percentage of easy VOIs that is incorrectly classified.

Table 3. Table with average percentages of all classified VOIs of which labels needed to be corrected for the different annotation protocols. In columns, from left to right, are results for easy VOIs, for which the observers agree on the label, hard VOIs, for which both observers did not agree on the label, and hard and easy VOIs taken together. All percentages for the interactive protocols are averaged over the ten times each experiment was performed and over all sixteen annotated datasets.

Protocol	Easy VOIs	Hard VOIs	All VOIs
Automatic (1)	55%	68%	60%
Interactive (2)	16%	31%	21%
Interactive with training data (3)	15%	28%	20%

4.3 Sensitivity and specificity

Table 4 shows the average sensitivities and specificities for all texture categories in the automatic and both interactive protocols. Results are split in results for the first slice, the rest of the scan after training on at least 50% of the VOIs in a scan and for all VOIs. In the automatic protocol, sensitivity ranges from 3% for nodular pattern to 66% for normal tissue. Specificity was lowest for normal tissue (76%). Specificity for all other textures was between 86% (ground glass) and 97% (decreased density). Comparable percentages can be seen after interactive classification of the first slice with previous training data. After interactive training on at least 50% of the VOIs in the interactive protocol without training data (protocol 2), sensitivities ranged between 43% for consolidation and 94% for nodular pattern. Specificity ranged between 94% for normal tissue and 100% for consolidation. Average percentages for classification of the rest of the scan in protocol 3 were similar. Results for all VOIs are similar to the results for the rest of the scan.

5 Discussion

In this study, we investigated different methods of automatic and interactive classification of VOIs containing normal lung tissue, decreased density, consolidation, honeycombing, ground glass, crazy paving, NSIP pattern and nodular pattern in thoracic CT scans of ILD patients. We compared interactive with automatic annotation, and we investigated the use of training data from previously annotated scans for classification of the first slice. The different protocols were tested using software we developed for simulation of interactive annotation. In previous work, we reported that interactive annotation of all VOIs in a limited number of scans was on average 3.7 times as fast as manual annotation of all VOIs (Kockelkorn et al. 2010). In this work, we evaluate three different (semi-)automatic classification approaches in terms of user interaction required to obtain completely annotated scans. These annotations can then be used to quantify the spread and the nature of disease processes.

Table 4. Sensitivities and specificities in % per texture for the different annotation protocols. For the two interactive protocols, sensitivity and specificity are indicated for classification of the first slice, for classification of all remaining VOIs after training on at least 50% of all VOIs, and for all VOIs. For the automatic protocol, overall sensitivity and specificity are given. Percentages are averaged over all ten repetitions per experiment and over all annotated datasets. NT = normal tissue; DD = decreased density; CO = consolidation; HC = honeycombing; GG = ground glass; CP = crazy paving; NS = NSIP pattern; NO = nodular pattern.

			NT		DD		CO		HC		GG		CP		NS		NO	
			sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
automatic	❶	all	66%	76%	45%	97%	27%	96%	13%	94%	36%	86%	44%	92%	49%	93%	3%	96%
interactive	❷	first slice	100%	0%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%
		last 50%	91%	94%	88%	98%	43%	100%	75%	99%	76%	97%	90%	99%	75%	99%	94%	97%
		all	89%	84%	79%	99%	30%	100%	64%	99%	65%	96%	81%	99%	63%	98%	82%	97%
interactive with training data	❸	first slice	67%	80%	49%	97%	24%	96%	18%	94%	38%	86%	46%	91%	51%	92%	4%	96%
		last 50%	91%	93%	87%	99%	42%	100%	76%	99%	75%	97%	91%	99%	75%	99%	94%	97%
		all	87%	91%	83%	98%	31%	100%	65%	99%	69%	95%	85%	98%	67%	98%	83%	97%

For all annotation strategies, we assessed classification accuracy for all texture classes taken together. Higher annotation accuracy means that less user interaction is required in order to obtain a complete annotation of the lung tissue. When considering annotation of a complete scan, interactive annotation yielded on average a higher classification accuracy (79%-80%) than automatic classification (42%). A drawback in using an automatic classification approach is that thoracic CT scans are made using a variety of scanning parameters. In addition, different observers each have their own way of annotating scans, which results in high inter-observer variability (Sluimer et al. 2006, Uppaluri et al. 1999). In this study, we found an interobserver agreement of 63%, which supports the claim that annotation of lung textures is not trivial. VOIs for which the observers did not agree on the label were more likely to be classified incorrectly in all three protocols. Using training data from different observers and from CT scans obtained with different scanning parameters leads to an automatic classification system that is not able to classify VOIs with high accuracy. In addition, high inter-observer variability makes it difficult to define a gold standard for VOI classification.

In this study, we did not aim for obtaining consensus ground truth annotations, or for selecting scans with similar acquisition parameters. Instead, we approached the problem of computer-aided annotation from the other side. We considered both the presence of interobserver variability and the lack of a uniform scanning protocol as facts we have to accept when working with clinical data. And whereas these facts may pose problems in automatic classification methods, our interactive method is able to overcome both. By using observer input -in this case corrections and approvals of automatically assigned labels- as

training data for classification of unseen VOIs in the scan, interactive annotation is able to adapt to the annotation style of the observer. Since the pool of training VOIs and classified VOIs are derived from the same scan, acquisition parameters are always the same. The success of this approach can be seen in Figure 7. Panels c and d confirm that different observers may disagree on the way textures should be annotated. When using training data from observers with different opinions, the resulting classifier will be unable to classify VOIs in a new scan with high accuracy (panel b). However, the interactive classification system learns the specifics of the annotation task from its observer and can therefore adapt to individual annotation requirements. This makes it a versatile system: It can adapt to different definitions of texture categories. In addition, each observer can define her or his own texture categories, if desired.

For any interactive system, user-friendliness is of pivotal importance. We have taken several measures to increase the user-friendliness of our interactive approach. First, we chose to divide the lungs into VOIs to simplify the task of delineating normal and abnormal areas in the lungs. These VOIs are constructed to capture one type of texture. In this way, the labor-intensive 3D annotation task is simplified to a number labeling of predefined VOIs.

Second, interactive classification starts with an untrained classifier. We compared two labeling methods for the first slice of a scan, both of which do not require any user interaction. If training data from previously annotated scans is available, this can be used for classification of the first slice. If no previous training data is present, a heuristic approach is to label all VOIs in that slice as normal tissue, yielding an average accuracy of 36%. A set up in which only training data from the same observer was used, could possibly yield even better predictions in classification of the first axial slice. However, more annotations per observer are needed to test this hypothesis.

Third, annotation takes place in a slice-based manner. The observer is shown axial slices with classification results, which he or she has to correct. Slices are chosen in such a way that all areas of the scan are visited during the training process. Within each area, the slice from which the classifier can learn most from the corrections of the observer is chosen.

The current experiment has some limitations. First, scans were selected in such a way that each texture that had to be discriminated was present in multiple scans. However, this did not result in a balanced dataset. Therefore, we balanced the previously obtained dataset by using 500 random samples per category for training. The interactively obtained datasets remain unbalanced. This should be taken into account when analyzing the overall classification accuracies. If a texture with a low number of instances, such as honeycombing, is not detected, this has a low impact on overall accuracy, while the presence of honeycomb-

ing is highly relevant in clinical practice. Normal tissue is the largest category. Hypothetically, we could devise a classifier that classifies all VOIs as normal tissue. This would result in an overall accuracy of 36%. Given that interactive classification yielded a considerably higher accuracy and that the distribution of texture in an unseen scan will very likely also be imbalanced as well, we do not consider this imbalance to be a major problem. In addition, the dataset that was used for the experiments contained a small number of scans, which may lead to a bias in the results. Therefore, repetition of the experiments using a larger dataset is desirable.

Second, all experiments involving interactive annotation in this research were simulated, instead of performed by human observers. Given the high inter-observer variability measured in this and other works, it is conceivable that repeating these experiments with human observers would lead to different results. We nevertheless chose to follow this simulation approach because it allowed us to repeat each experiment ten times, using different parameters, without requiring annotation effort from radiologists.

Third, we have only performed a limited search to find the optimal feature set for our (semi-)automatic classification protocols. We looked into the use of Haralick features in the automatic classification experiments, to see whether these performed better than the set used in this work. This led to an average accuracy of 38%. In addition, we investigated the relative feature importance, as reported by the Extra-Trees classifier, averaged over all annotated datasets and all 10 repetitions of experiments, of the original feature set used in this study for all three protocols. We compared the fifteen most important features in each protocol, and twelve of these were selected in all three scenarios. These were: mean, skew, and kurtosis after Gaussian filtering at scale σ=1, mean and skew after Gaussian filtering at scale σ=2, mean after Gaussian filtering at scale σ=4, mean and standard deviation of the second Hessian eigenvalue at scale σ=2, mean, standard deviation, and kurtosis of the third Hessian eigenvalue at scale σ=1 and standard deviation of the third Hessian eigenvalue at scale σ=2. We used this selected feature set to determine the overall percentage of label changes needed for complete annotation in scenarios 1, 2 and 3. Automatic classification required changing the labels of 59% of the VOIs (versus 58% using the original feature set). The overall percentages of the interactive protocols without and with previous data were 21%, versus 21% and 20% respectively. While these limited sets of experiments did not result in higher classification accuracy, a more elaborate comparison of feature sets and also classifiers may be beneficial for further development of the proposed methods.

Finally, we have shown that interactive annotation results converge to the preferences of the observer. The resulting annotations may not be universally accepted, for example if

the method is used by an inexperienced observer. In this respect, results should be treated as manual annotations: the reliability depends on the person performing the labeling.

In the future, we plan to test this interactive annotation tool more extensively with live observers. With the resulting annotations, an annotated ILD texture dataset can be built with a relatively small effort. This dataset can then be used for the training of radiologists, as a reference standard in clinical practice or for developing an automated classification tool. As a next step, interactive or automatic classification results may be beneficial in disease progression monitoring and for making the right diagnosis within the ILD spectrum.

6 Conclusion

We have presented one automatic and two interactive protocols for classification of VOIs in the lungs in thoracic CT scans of ILD patients. These protocols were compared in terms of user actions required to obtain complete and verified annotations. In both interactive protocols, VOIs were annotated in a slice-wise manner. For classification of the first slice, the interactive system benefited from the addition of training data from other scans, since the interactive classifier was then still untrained. In classification of the following slices, the interactive classifier learned to discern the different texture categories by asking the observer to correct misclassifications. By choosing the slices that were shown to the observer based on cumulative classifier uncertainty, annotation effort could be minimized. Using the optimal settings, the average accuracy for classification of the first slice was 44%. After interactive training on 50% of the VOIs in a scan, the remaining VOIs are classified with an average accuracy of 87%. Therefore, the interactive approach we describe is a promising tool to make annotation of complete scans practically feasible.

References

Aziz ZA, Wells AU, Bateman ED, Copley SJ, Desai SR, Grutters JC, Milne DG, Phillips GD, Smallwood, D, Wiggins J, Wilsher ML, Hansell, DM. Interstitial lung disease: effects of thin-section CT on clinical decision making. *Radiology.* **2006;**238(2):725–733.

Depeursinge A, Iavindrasana J, Hidki A, Cohen G, Geissbuhler A, Platon A, Poletti P-A, Müller H. Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization., *Journal of digital imaging.* **2010;** 23(1):18–30.

Depeursinge A, Van de Ville D, Platon A, Geissbuhler A, Poletti P-A, Müller H. Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames. *IEEE Transactions on Information Technology in Biomedicine.* **2012**;16(4):665–675.

Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Machine Learning.* **2006**;63(1):3–42.

Hu S, Hoffman E, Reinhardt J. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE Transactions on Medical Imaging.* **2001**;20:490–498.

Huber MB, Bunte K, Nagarajan MB, Biehl M, Ray LA, Wismüller A. Texture feature ranking with relevance learning to classify interstitial lung disease patterns., *Artificial Intelligence in Medicine.* **2012**; 56(2):91–97.

Huber MB, Nagarajan MB, Leinsinger G, Eibel R, Ray LA, Wismüller A. Performance of topological texture features to classify fibrotic interstitial lung disease patterns., *Medical Physics.* **2011**;38(4): 2035–2044.

Kockelkorn TTJP, Schaefer-Prokop CM, Bozovic G, Muñoz-Barrutia A, van Rikxoort EM, Brown MS, de Jong PA, Viergever MA, van Ginneken B. Interactive lung segmentation in abnormal human and animal chest CT scans, *Medical Physics.* **2014**;41(8): 081915.

Kockelkorn TTJP, de Jong PA, Gietema HA, Grutters JC, Prokop M, van Ginneken, B. Interactive annotation of textures in thoracic CT scans, *Proceedings of the SPIE* **2010**; 7624:76240X–76240X–8.

Park SO, Seo JB, Kim N, Park SH, Lee YK, Park B-W, Sung YS, Lee Y, Lee J, Kang S-H. Feasibility of automated quantification of regional disease patterns depicted on high-resolution computed tomography in patients with various diffuse lung diseases. *Korean J Radiol.* **2009**;10(5):455–463.

Settles B. Active learning literature survey. *Computer Sciences Technical Report.* **2009**;1648. University of Wisconsin–Madison.

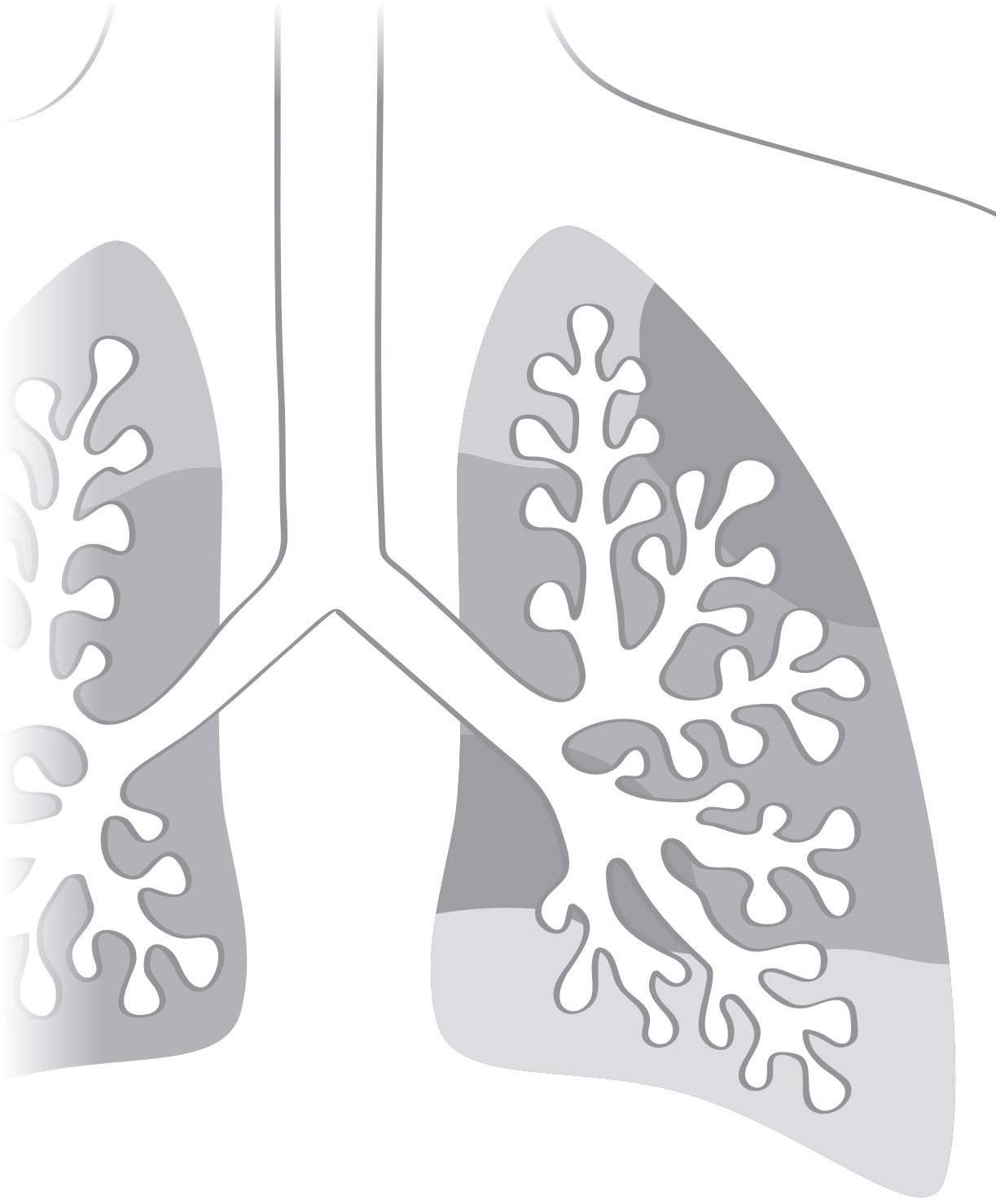
Sluimer IC, Prokop M, Hartmann I, van Ginneken B. Automated classification of hyperlucency, fibrosis, ground glass, solid and focal lesions in high resolution CT of the lung. *Medical Physics.* **2006**;33(7):2610–2620.

Uppaluri R, Hoffman EA, Sonka M, Hunninghake GW, McLennan G. Interstitial lung disease: A quantitative study using the adaptive multiple feature method. *Am J Respir Crit Care Med.* **1999**;159(2):519-25.

van Rikxoort EM, de Hoop B, Viergever MA, Prokop M, van Ginneken B. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection, *Medical Physics.* **2009**;36(7):2934–2947.

Xu Y, van Beek EJ, Hwanjo Y, Guo J, McLennan G, Hoffman EA. Computer-aided classification of interstitial lung diseases via MDCT: 3D adaptive multiple feature method (3D AMFM). *Acad Radiol.* **2006**;13(8):969-78.

Zavaletta VA, Bartholmai BJ, Robb RA. High resolution multidetector CT-aided tissue analysis and quantification of lung fibrosis. *Acad Radiol.* **2007**;14(7):772-87.



Optimization strategies for interactive classification of interstitial lung disease textures

Thessa TJP Kockelkorn
Rui Ramos
José Ramos
Pim A de Jong
Cornelia M Schaefer-Prokop
Rianne Wittenberg
Audrey M Tiehuis
Jan C Grutters
Max A Viergever
Bram van Ginneken

Conditionally accepted for publication in Frontiers in ICT

Abstract

For computerized analysis of textures in CT scans of interstitial lung disease patients, manual annotations of lung tissue are necessary. Since making these annotations is labor-intensive, we previously proposed an interactive annotation framework. In this framework, observers iteratively trained a classifier to distinguish the different texture types by correcting its classification errors. In this work, we investigated three ways to extend this approach, in order to decrease the amount of user interaction required to annotate all lung tissue in a CT scan. First, we conducted automatic classification experiments to test how data from previously annotated scans can be used for classification of the scan under consideration. We compared the performance of a classifier trained on data from one observer, a classifier trained on data from multiple observers, a classifier trained on consensus training data, and an ensemble of classifiers, each trained on data from different sources. Experiments were conducted without and with texture selection. In the former case, training data from all eight textures was used. In the latter, only training data from the texture types present in the scan were used, and the observer would have to indicate textures contained in the scan at hand. Second, we simulated interactive annotation to test the effects of (1) asking observers to perform texture selection before the start of annotation, (2) the use of a classifier trained on data from previously annotated scans at the start of annotation, when the interactive classifier is untrained, and (3) allowing observers to choose which interactive or automatic classification results they wanted to correct. Finally, various strategies for selecting the classification results that were presented to the observer were considered. Classification accuracies for all possible interactive annotation scenarios were compared. Using the best-performing protocol, in which observers select the textures that should be distinguished in the scan and in which they can choose which classification results to use for correction, a median accuracy of 88% was reached. The results obtained using this protocol were significantly better than results obtained with other interactive or automatic classification protocols.

1 Introduction

In medical image analysis, obtaining reliable ground truth annotations is of pivotal importance for the training, testing, and comparison of algorithms. Several factors may hamper the construction of such a ground truth dataset. To start, making annotations is expensive, as it requires a substantial amount of human observer effort. But before observers can start annotating, a selection of scans needs to be made. This selection process has a large influence on the quality and usability of the resulting dataset. When choosing images from

a single institution, with similar acquisition and reconstruction parameters, training and testing are done on a homogeneous dataset. Good performance of an algorithm on such a dataset is by no means a guarantee of comparable results on other datasets. However, when choosing to collect a set of images with varying acquisition parameters, training becomes more complicated. In this case, it is likely that more training data is needed to obtain results similar to the ones that can be obtained using a homogeneous dataset. Finally, further complications arise when the ground truth is difficult to establish. The presence or absence of a bone fracture will in most cases not give rise to much debate, since most bone fractures can be established objectively. For more subtle lesions, matters readily become more complicated, since different observers may have different opinions on the interpretation of these lesions.

In this work we consider interstitial lung disease (ILD), a group of around 200 inflammatory and fibrotic lung diseases that mainly affect the tissue and space around the air sacs of the lungs. These diseases have distinct, but also considerably overlapping imaging features. Since the individual diseases have substantially different treatment options and prognosis, it is important to make the correct diagnosis. Computed tomography (CT) scans play an important role in the interdisciplinary process of making a diagnosis of ILD (Aziz et al. 2006). Automatic classification of normal and abnormal lung tissue in CT scans of ILD patients has been studied extensively, focusing on finding optimal feature sets and classifiers (Uppaluri et al. 1999, Xu et al. 2006, Zavaletta et al. 2007, Sang et al. 2009, Depeursinge et al. 2010, Depeursinge, Van De Ville, et al. 2012, Huber et al. 2011, Huber et al. 2012, Vasconcelos et al. 2015) Recently, deep learning techniques have been applied to the problem (Gao et al. 2015, Anthimopoulos et al. 2016)

The focus of this study is on the process of obtaining annotations which can be used in ILD texture analysis. For scans from patients with ILD, all complicating aspects mentioned above may occur when compiling an annotated dataset. First, manual delineation of all lung textures present in a volumetric CT scan is a labor-intensive task, especially when the disease affected a large part of the lungs. Second, thoracic CT scans of ILD patients may be made using various CT protocols (Prosch et al. 2013). Finally, analysis of imaging features and therefore also annotations vary substantially . Therefore, obtaining ground truth annotations is not trivial.

Since automatic ILD annotation systems may not be able to adapt to a wide variety of CT acquisition protocols, let alone to different annotation preferences, we have developed a system for interactive annotation of 3D volumes of interest (VOIs), which we applied to scans of ILD patients (Kockelkorn et al. 2010, Kockelkorn et al. 2016). This method al-

lows observers to quickly annotate ILD textures in CT scans. A human observer trains the system continuously by correcting classification results in a slice-by-slice manner. In this way, the algorithm becomes increasingly better in annotating textures. The smaller the amount of user input required to obtain completely annotated lungs, the easier it becomes to obtain a large number of annotated datasets, which can then for example be used to study effects of varying acquisition parameters on (semi-)automatic classification accuracy.

In the current work, we aim to optimize the interactive annotation procedure, in order to decrease the numbers of VOIs for which the observer has to correct the computer-generated label. We investigate how annotations of VOIs in other previously annotated ILD scans can be used for classification of unseen VOIs in the scan under consideration. In addition, we evaluate various ways in which users transfer their knowledge to the interactive annotation environment. Finally, we compared different strategies for selection of the VOIs that are shown to the observer for correction.

This paper is structured as follows: Section 2.1 describes the CT scans used for the experiments. Section 2.2 details the processes of automatic and interactive classification, followed by an outline of the experiments that were performed in Section 2.3. Section 3 contains the results of the experiments. In Section 4 the main insights resulting from this work are summarized and their relevance is discussed.

2 Material and Methods

2.1 Materials

For this project, 23 clinically indicated, standard-dose thoracic CT scans of ILD patients were collected retrospectively. Scans were acquired between 2004 and 2010 at the St Antonius Ziekenhuis Nieuwegein, the Netherlands, on a Philips Mx8000 IDT or a Philips Brilliance iCT scanner (Philips Medical Systems, Best, the Netherlands). Scans were taken at full inspiration with patients in supine position, without contrast material. Data were acquired in spiral mode and reconstructed to 512×512 or 768×768 matrices. Patient and scanning protocol parameters are summarized in Table 1.

2.2 Methods

2.2.1 Preprocessing

Interactive annotation of VOIs in the lungs has been described previously (Kockelkorn et al. 2016). To summarize, lungs in the CT scans were segmented (van Rikxoort et al. 2009)

Table 1. Patient and scan characteristics.

Scan nr	Patient age (y)	Patient sex	Nr of VOIs	In-plane resolution (mm)	Slice spacing (mm)	Peak voltage (kV)	Tube current (mA)
1	67	female	2021	0.605	0.8	120	120
2	20	female	2786	0.574	0.8	120	217
3	33	male	3084	0.688	0.8	120	144
4	45	male	1647	0.873	0.8	120	144
5	42	female	2040	0.658	0.8	120	144
6	61	male	1234	0.781	1.0	120	192
7	24	male	2818	0.645	0.7	120	206
8	32	male	1459	0.758	0.8	120	144
9	62	female	1368	0.688	0.5	120	188
10	38	female	2386	0.586	1.0	120	270
11	57	female	1148	0.660	0.8	120	90
12	71	male	1761	0.707	0.8	120	144
13	51	male	2017	0.652	0.8	120	144
14	57	male	1692	0.750	0.8	120	90
15	33	male	1701	0.666	0.8	120	90
16	52	male	1810	0.676	0.5	120	188
17	30	male	3177	0.658	0.8	120	90
18	39	female	3313	0.627	0.5	120	188
19	58	male	1687	0.688	0.8	120	90
20	49	female	2233	0.580	0.8	120	150
21	54	female	1681	0.672	0.8	120	90
22	53	female	2540	0.411	4.0	120	125
23	32	female	2579	0.652	0.8	120	90

and subdivided into roughly spherical VOIs, containing only one type of texture, using the algorithm described in (Kockelkorn et al. 2016). On average, lungs contained 2114 VOIs (range: 1148-3313).

2.2.2 Features and classifier

For all VOIs, 72 rotationally invariant features were calculated. Scans were filtered using Gaussian, Laplacian, gradient magnitude, and three Hessian eigenvalue-based filters. Each of these six filters was applied at three scales (σ = 1, 2, and 4 voxels). This resulted in 18 filtered images. In each filtered image, mean, standard deviation, kurtosis, and skew of the VOI histogram were calculated and used as features. Because training of the classifier is done while the observer is annotating a scan, we used an Extra-Trees classifier (99 trees,

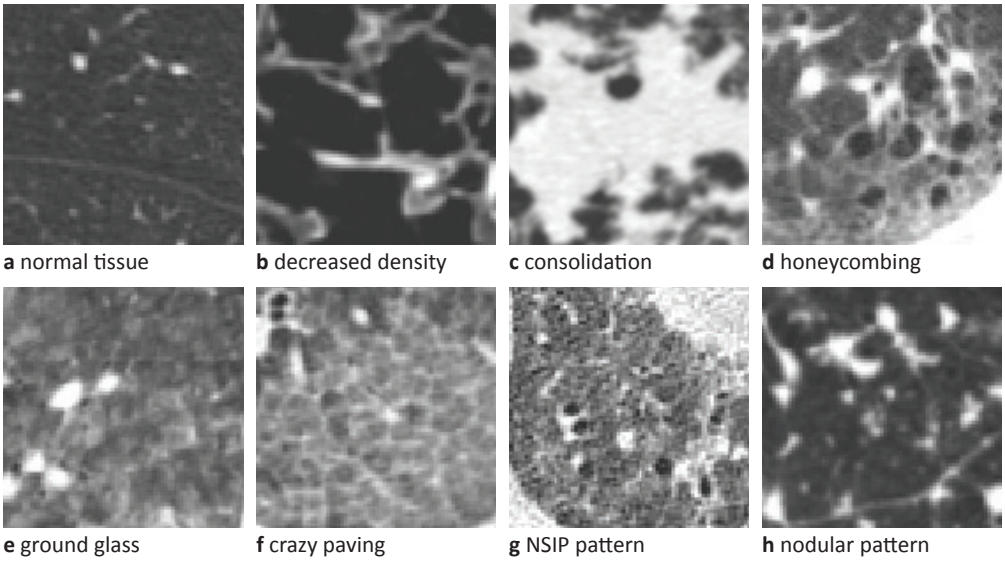


Figure 1. Examples of the 8 texture classes

10 features per node, a minimum number of samples after each split of 1; parameters based on the recommendations of Geurts et al. (2006)). This classifier can be trained relatively fast, since the training process can be divided over multiple cores.

2.2.3 Manual annotation of VOIs

Three radiologists independently performed manual labeling of all VOIs in a subset of the set of 23 scans. Two radiologists labeled 6 scans and one radiologist labeled 21 scans. In total, 17 scans were annotated by one observer, 2 were annotated by 2 observers, and 4 were annotated by all 3 observers.

Observers were instructed to annotate the following textures (see Figure 1):

Normal tissue Lung tissue without any abnormalities
Ground glass Increased lung density, in which underlying structures are still visible
Consolidation Increased lung density, in which underlying structures are no longer visible
Honeycombing Cystic destruction of subpleural lung parenchyma: there are cysts of varying diameter (0.3-1.0 cm) in several layers and cysts share relatively thick walls
Decreased density Decreased density compared with normal lung parenchyma, with or without surrounding walls
Crazy paving Regular pattern of ground glass with a reticular pattern

Table 2. Percentages of VOIs assigned to the 8 texture classes

Normal tissue	Decreased density	Consolidation	Honeycombing	Ground glass	Crazy paving	NSIP pattern	Nodular pattern
55%	15%	1%	2%	8%	6%	1%	11%

NSIP pattern Ground glass with architectural distortion, traction bronchiectasis or irregular lines

Nodular pattern Sharply defined nodular densities (1-4 mm) in a random or paralymphatic (paraseptal) distribution. Nodules can also have branching structures (tree-in-bud)

Thus, in this work an 8-class classification problem is studied. Table 2 shows the percentages of VOIs assigned to each of the 8 classes in the resulting dataset.

2.2.4 Simulated interactive annotation

We used simulation software to investigate the effect of design choices in interactive annotation protocols on the percentage of VOIs that was correctly classified. Interactive annotation is schematically depicted in Figure 2. Initially, an axial slice was chosen at random and VOIs intersecting with this slice were labeled automatically. Automatically generated labels were compared to the manual labels that were provided by the human observers and incorrect labels were changed. All labeled VOIs were used to train an Extra-Trees classifier. A second axial slice was chosen and all VOIs intersecting with the slice were classified by the classifier. Also in this slice, automatic classifications were compared to manual annotations and incorrect labels were changed. The VOIs in this second slice were added to the training dataset, on which the Extra-Trees classifier was retrained. The cycle of correction, retraining, and classification was repeated until at least 50% of all VOIs were annotated. The remainder of the scan was then classified automatically and subsequently checked against the manual annotations. The percentage of correctly classified VOIs was used as a performance indicator.

2.3 Experiments

To determine an optimal protocol for interactive annotation, three types of experiments were conducted. First, automatic classification experiments were done to evaluate which type of previous training data yielded the best classification results in unseen scans. Second, we evaluated different interactive annotation scenarios. We investigated several ways to exploit prior knowledge about the annotation task. Finally, experiments were conducted to compare four strategies for slice selection. As all experiments included random selection of training samples, they were repeated 5 times, after which results were averaged.

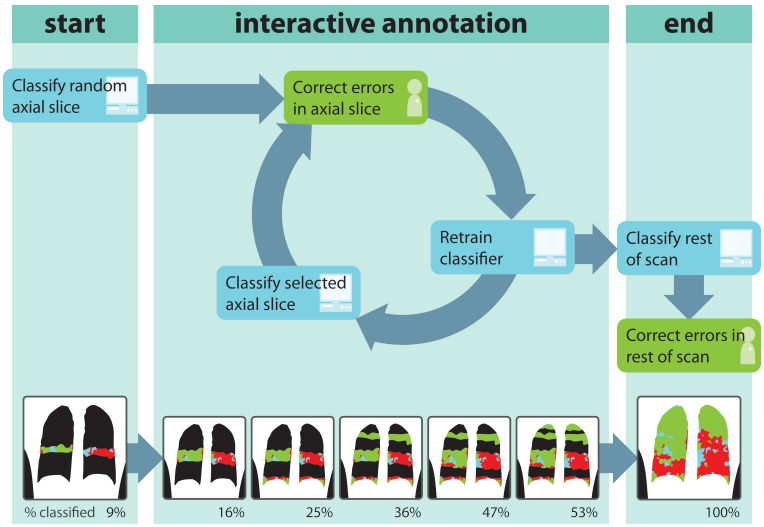


Figure 2. Flowchart of interactive VOI-based annotation. Green boxes are (simulated) user actions; blue boxes are computer actions. In the bottom of the figure, resulting annotations at for the different steps in the procedure are shown. The percentages underneath the images are the percentages of VOIs in the scan that have received a label.

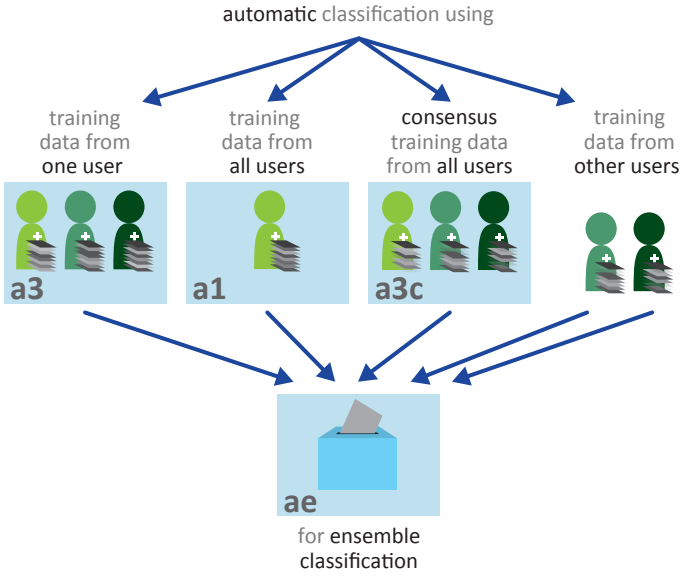


Figure 3. Schematic overview of the four automatic annotation protocols. In protocol a1 (left), training data from the observer whose annotations are used as ground truth for this dataset are used to train the classifier. In protocol a3 (second from the left), training data from all observers is used to train the classifier. In protocol a3c (second from the right), consensus annotations are used as training data. These consensus annotations are the VOIs that received the same label independently from 2 or 3 observers. In protocol ae (bottom), an ensemble of 5 classifiers was trained. Classifiers were trained on the following types of training data: data from the same observer (a1), data from each of the two other observers (right), data from all observers (a3) and from the consensus annotations of multiple observers (a3c). The final label was determined by voting of these 5 classifiers.

2.3.1 Automatic classification experiments

Four automatic classification protocols were tested in a leave-one-scan-out set-up (see Figure 3). In all automatic protocols, 100, 250 or 500 training VOIs were selected for each texture pattern. If a texture was represented by less than the intended number of samples, all samples were used. In the first approach (a1), we trained an Extra-Trees classifier on training data from other scans annotated by the same observer that annotated the scan under consideration. In the second approach (a3), the classifier was trained using training data from other scans annotated by all three observers. In the third approach, we trained a classifier on consensus training data (a3c). This training data set was obtained by selecting VOIs that were labeled two or three times. In the first case, the two observers had to agree on the label. In the second case, all three observers had to agree. If less than 100 consensus samples were available for a texture, non-consensus training samples from all users were used. In the fourth approach, we used an ensemble classification strategy. For each dataset annotated by observer x, Extra-Trees classifiers were trained on different types of training data:

I training data from other scans annotated by observer x
II training data from other scans annotated by all observers
III consensus training data from other scans annotated by all observers
IV training data from other scans annotated by observer y
V training data from other scans annotated by observer z

The final class label was determined by voting. In case of a tie, the class with the highest posterior probability was chosen. Results were obtained for fully automatic classification of all annotated datasets.

Experiments with approaches a1, a3, a3c, and ae were performed with and without texture selection. Without texture selection, training data from all 8 texture types was used to train the classifier. With texture selection, only training data from the texture types present in the scan was used (see section 2.3.2.1).

2.3.2 Interactive classification experiments

Interactive classification was simulated in a leave-one-scan-out set-up, as described in (Kockelkorn et al. 2016). We investigated three manners to decrease annotation effort.

2.3.2.1 Texture selection

First, we studied the effect on classification accuracy of indicating 5 VOIs of each texture present in the scan before the start of annotation in the different interactive protocols.

This approach is illustrated in Figure 4a. The major difference in these protocols and the ones in Figure 4b was the way in which the VOIs in the first axial slice were classified. In the left protocol (i-ts), classification was performed interactively by a classifier trained on the VOIs selected by the simulated observer before training of the interactive classifier. In the protocols using automatic classification results for the classification of the first slice (i-a1, i-a3, i-a3c, and i-ae), only training samples from the texture categories indicated by the user before annotation were used. Obviously the classification problem is simplified considerably if the system knows from the start which of the 8 texture classes are present in the scan.

2.3.2.2 Training data from previously annotated scans

Second, we investigated the effect of using classifiers trained on training data obtained from previously annotated scans. These automatic classification results could be used at the beginning of interactive annotation, when little or no training data from the scan under consideration is available.

We tested the following annotation protocols, schematically depicted in Figure 4a (with texture selection) and Figure 4b (without texture selection):

i: completely interactive annotation, without the use of previous training data
i-a1, i-a3, i-a3c and i-ae: classification of VOIs in the first slice using one of the four automatic classification methods described above, followed by interactive annotation
i-cc: annotation in which observers could determine per slice which classification results they wanted to use as a starting point for corrections - this could be i, a1, a3, a3c or ae.

2.3.2.3 Classifier choice

Third, we investigated the scenario in which observers were given the option to choose from different classification results when correcting the labels of the individual VOIs. The protocols based on this approach are i-cc-ts in Figure 4a and i-cc in Figure 4b. Users could choose between interactive classification results, and results of a classifier trained on data from other scans annotated by the observers themselves (a1), a classifier trained on data from other scans annotated by all observers (a3), a classifier trained on consensus training data from other scans (a3c), and the ensemble classification method (ae). For correction of the first axial slice in the absence of previous training data, without texture selection, the user could choose for heuristic labeling instead of interactive labeling. From these different classification results, the one requiring the lowest number of corrections was selected in the simulations.

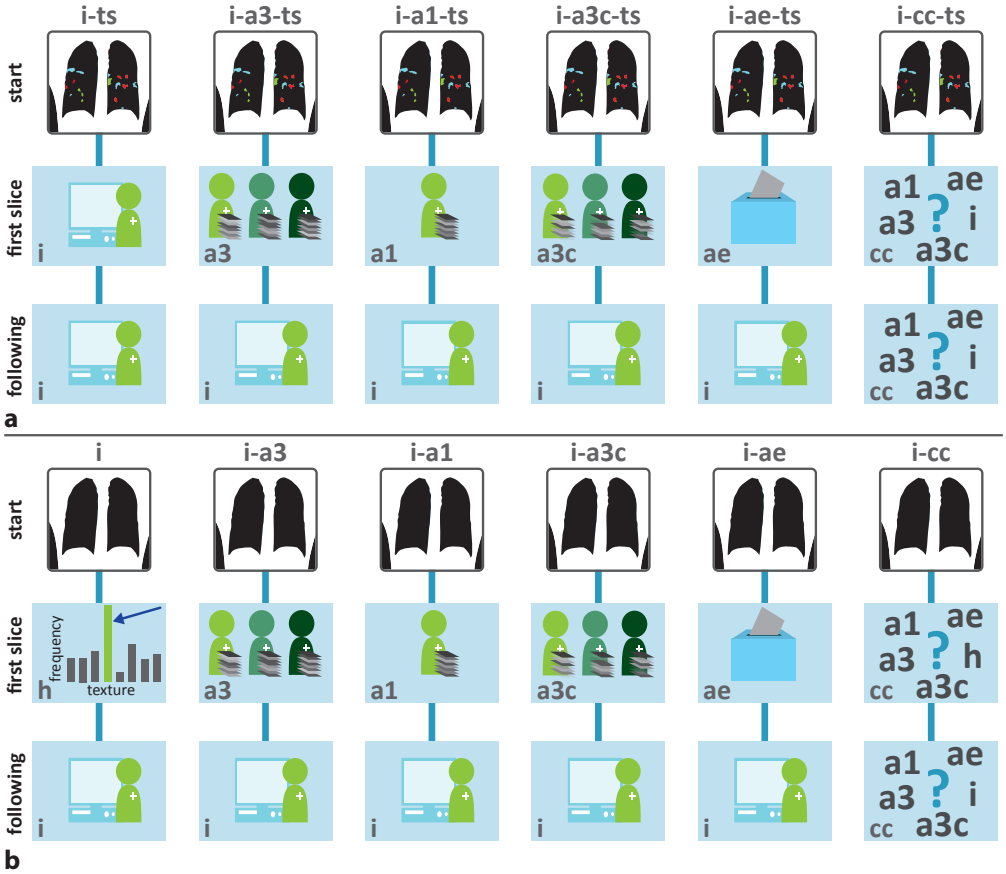


Figure 4. a. Schematic overview of the tested interactive annotation protocols with texture selection (ts). The top row displays how classification is initiated: the user selected 5 example VOIs of each texture class present in the scan. The images in the second row indicate the classification procedure for the VOIs in the first slice. This can be done interactively, by training the classifier on the VOIs that the user indicated before the start of annotation (i: interactive), by using a classifier trained on the automatic classification protocols in Figure 3 (a1, a3, a3c, and ae), or by letting observers choose the classification results that they want to correct (cc: classifier choice). The third row indicates how the following slices are classified: either interactively (i) or by letting observers choose the classification results that they want to correct (cc). **b.** Schematic overview of the tested interactive annotation protocols without texture selection. The top row displays how classification is initiated: in this case no user action is required at this stage. The second and third rows are similar to panel a.

2.3.2.4 Slice selection

Finally, the way in which slices are presented to the observer influences the efficiency of the training of the classifier. We compared random slice selection with selection based on cumulative uncertainty U of n unassigned VOIs per slice:

$$U = \sum_{v=1}^{n} (1 - c_v) \qquad (1)$$

In this formula, the uncertainty of the classifier for VOI v was calculated by subtracting the confidence c_v from one. c_v was the highest posterior probability among the classes. By adding the uncertainties for all unassigned VOIs in a slice, its cumulative uncertainty was determined. In each classification and correction cycle, the slice with the highest cumulative uncertainty was chosen. We chose to use cumulative instead of average uncertainty, since the first method favored slices with large numbers of VOIs. Larger numbers of VOIs per slice meant that the cycle of retraining, classification, and correction had to be repeated a smaller number of times.

In our previous study, we subdivided the lungs into 5 levels in axial direction, numbered 1-5 from the apex to the base of the lungs (Kockelkorn et al. 2016). Slices are chosen from levels 3 - 5 - 2 - 4 - 1 - 3, and so on. In this work, we compared random and uncertainty-based slice selection in a scenario in which this subdivision was used. In addition, we tested the effects of random and uncertainty-based slice selection when the lungs were not subdivided and slices could be chosen from the apex to the base of the lungs in each classification and correction cycle. This resulted in the following slice selection methods:

I random selection from the entire lungs
II uncertainty-based selection from the entire lungs
III random selection from different levels in the lungs
IV uncertainty-based selection from different levels in the lungs

2.3.3 Evaluation

In all automatic experiments, classification accuracy was calculated per annotated dataset. For the interactive classification experiments, classification accuracy was calculated for each classified slice, for classification of the remainder of the scan after training on at least 50% of all VOIs, and for the complete dataset. In all cases, results for the five repetitions per experiment were averaged per annotated dataset. If a scan was annotated by more than one observer, results were calculated for all 2 or 3 annotations separately.

Repeated measures ANOVA was performed to test the difference in overall accuracy between the interactive protocols without texture selection (i, i-a1, i-a3, i-a3, i-ae, and i-cc), between the best-performing interactive protocol with and without texture selection, and between the best performing interactive protocol with texture selection and the four automatic protocols (a1, a3, a3c, and ae) with texture selection.

Table 3. Interobserver agreement in % for VOIs that were annotated two or three times.

		% of VOIs that received		
	Nr of VOIs	**3 labels**	**2 labels**	**1 label**
2 times annotated	4724	-	31	69
3 times annotated	8498	30	35	35

3 Results

3.1 Interobserver agreement

Four scans in our dataset were annotated by 3 observers and 2 scans were annotated by 2 observers. Table 3 shows the results of the comparison of the labels of the VOIs that were annotated more than once. 4724 VOIs were annotated twice. Interobserver agreement was 69%. 8498 VOIs were annotated three times. 35% of the VOIs were given the same label by all three observers. For another 35% of the VOIs, 2 observers agreed on the label and one observer had a different opinion. The final 30% of the VOIs received a different label from each of the three observers.

Figure 5 shows an example of an axial slice (a), with in the second row manual VOI annotations as made by observer 1 (d), observer 2 (e), and observer 3 (f). All observers agree that this slice contains normal tissue and crazy paving, but the distribution of the textures varies between the observers. In addition, observer 2 has indicated areas of ground glass, which are absent in the annotations of observers 1 and 3.

3.2 Automatic classification results

3.2.1 Use of previous training data

In Figure 6a, the results of the use of different protocols for automatic classification -namely using training data from the observer that provided annotations for the scan under consideration (a1), using training data from all observers (a3), using consensus training data (a3c) and building an ensemble classifier (ae)- are displayed. The lower border of each box is the first quartile (Q1); the upper border is the third quartile (Q3). The median value is indicated by the horizontal bar inside the box. Upper and lower whiskers extend to the maximum and minimum data point, respectively. Results are shown for three different intended sizes of the training dataset: 100, 250, and 500 training samples per texture class. For all four classification protocols, adding more training samples yielded higher median classification accuracy. This also held for the maximum accuracy and for Q3. The minimum accuracy was 0% or close to 0% for all different protocols and for the three dif-

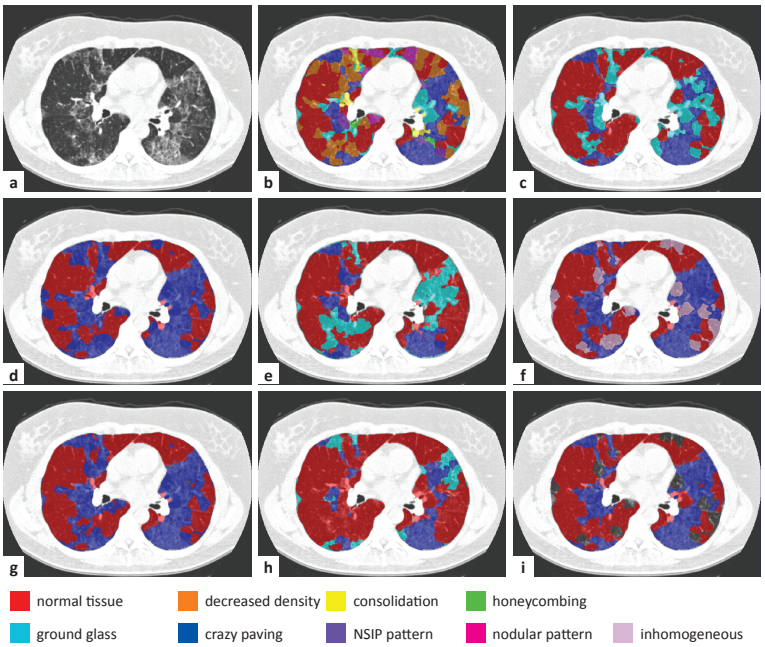


Figure 5. Examples of automatic and interactive classification results. **a.** Axial slice of a CT scan of an ILD patient. **b.** Results of automatic classification using training data from all observers without texture selection. **c.** Results of automatic classification with texture selection. **d.** Manual VOI labeling by observer 1. **e.** Manual VOI labeling by observer 2. **f.** Manual VOI labeling by observer 3. **g.** Interactive classification results after training on at least 50% of all VOIs for observer 1. **h.** Interactive classification results after training on at least 50% of all VOIs for observer 2. **i.** Interactive classification results after training on at least 50% of all VOIs for observer 3.

ferent intended training dataset sizes. The interquartile range (IQR), calculated as Q3-Q1, increased with an increasing intended number of training samples per texture class for the protocols a1 and a3c.

Table 4 shows the median, minimum, and maximum accuracies for all four automatic classification protocols, for each of the three observers, and for all observers taken together. The intended number of training samples per texture class was 500. The protocol in which consensus training data was used yielded the highest median percentage of correctly classified VOIs for observer 1, observer 3, and for all observers together. The protocol yielding the highest median accuracy for observer 2 was a1, in which only training data from this observer was used. The minimum percentage of correctly classified VOIs varied between the different observers: for observer 1, this percentage varied from 22% to 35%. For observer 3, values ranged between 8% for protocol a1 and 18% for protocol a3c. Minimum values were lowest for observer 2.

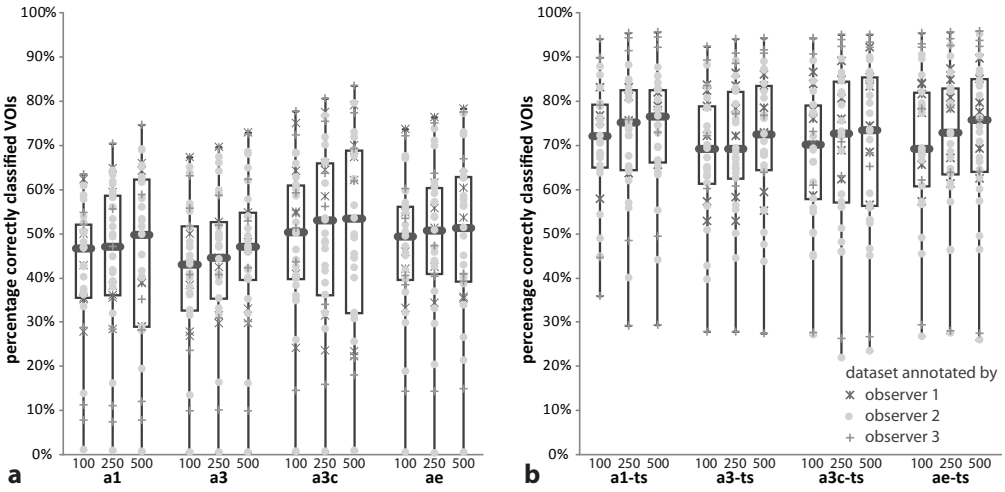


Figure 6. Box plots displaying classification accuracies of all 33 annotated datasets for different automatic scenarios. **a**. Results for automatic classification without texture selection. **b**. Results for automatic classification with texture selection. Each box displays the median value (horizontal thick line), the first quartile (Q1, lower margin of the box), and the third quartile (Q3, upper margin of the box). The lower whisker extends to the smallest data point; the upper whisker extends to the largest data point. a1: training data from same user; a3: training data from all users; a3c: consensus training data; ae: ensemble classification; ts: texture selection.

Table 4. Median (and range) of percentages of correctly classified VOIs for observer 1, observer 2, observer3, and for all 33 annotated datasets in the four automatic annotation protocols without texture selection. For each texture class, the intended number of training samples was 500.

	Median (min-max) % of correctly classified VOIs per protocol			
Protocol	**Obs 1**	**Obs 2**	**Obs 3**	**All**
One user (a1)	39 (29-66)	53 (1-69)	32 (8-75)	50 (1-75)
Three users (a3)	47 (30-73)	47 (0-68)	48 (10-72)	47 (0-73)
Consensus (a3c)	65 (22-79)	49 (1-79)	66 (18-83)	53 (1-83)
Ensemble (ae)	57 (35-78)	51 (0-77)	41 (15-77)	51 (0-78)

3.2.2 Texture selection

Figure 6b shows the effect of texture selection on classification accuracy for the 4 automatic protocols. When using texture selection, median classification accuracy substantially increased for all protocols as compared to the results without texture selection. Minimum and maximum accuracy also showed this increase. Similar to the situation in which no texture selection was applied, median classification accuracy increased with increasing size of the training dataset.

In Table 5, median, minimum, and maximum classification accuracy after texture selection are shown for the individual observers and for all observers together. The spread in medi-

Table 5. Median (and range) of percentages of correctly classified VOIs for observer 1, observer 2, observer3, and all 33 annotated datasets in the four automatic annotation protocols with texture selection. For each texture class, the intended number of training samples was 500. Results are taken over all annotated datasets.

	Median (min-max) % of correctly classified VOIs per protocol			
Protocol	**Obs 1**	**Obs 2**	**Obs 3**	**All**
One user (a1)	78 (66-82)	74 (44-88)	83 (29-96)	77 (29-96)
Three users (a3)	76 (55-86)	69 (44-89)	84 (27-94)	73 (27-94)
Consensus (a3c)	79 (56-92)	73 (23-90)	80 (27-95)	73 (23-95)
Ensemble (ae)	78 (62-90)	73 (26-95)	84 (27-96)	76 (26-96)

an accuracies between the different protocols was smaller than when no texture selection was applied. For all observers, median accuracy ranged from 73% for protocols a3c and a3, to 77% for protocol a1. This latter was an increase of 27 percentage points as compared with the scenario in which no texture selection was performed. The same could be seen for all individual observers: the spread in median accuracy between the classification protocols was also smaller when texture selection was applied.

Panels b and c in Figure 5 show the effect of texture selection. In both panels, VOIs were automatically classified using training data from all observers. In Panel b, this was done without texture selection, in Panel c with texture selection. Without texture selection, the classifier predicted the occurrence of all texture classes in this slice; with texture selection, the areas containing normal tissue as indicated by all observers are more accurately classified.

3.3 Interactive classification results

3.3.1 Texture selection

The following paragraphs contain the results of interactive annotation experiments. First, we studied the effect of texture selection in interactive classification. In 20 of the 33 annotated datasets, two types of textures were found. In 9 scans 3 textures were selected, and in the remaining 4 scans, 4 textures were selected. In Figure 7, classification accuracies per slice and for the remainder of the scan after training on at least 50% of the VOIs in the scan are shown for interactive protocols i (without texture selection) and i-ts (with texture selection). In both protocols, no previously annotated training data were used. In the first axial slice, median classification accuracy was the same whether VOIs were all labeled as normal tissue (in i) or VOIs were classified using a training dataset of 5 samples per texture class present in the scan. However, the IQR and the total accuracy range were larger if no texture selection was performed. The main advantage of texture selection in

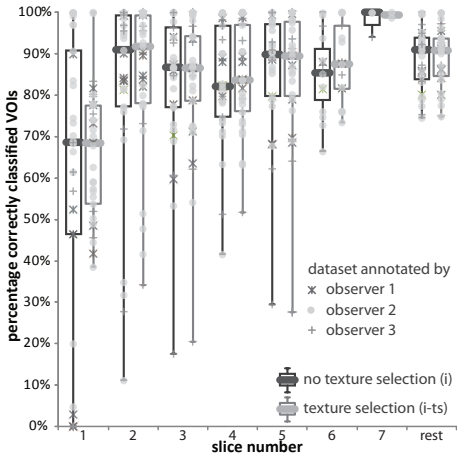


Figure 7. Box plot displaying the classification accuracies per slice and for the remainder of the scan for all 33 annotated datasets for interactive protocols i (darker grey boxes) and i-ts (lighter grey boxes). See the caption of Figure 6 for an explanation of the boxes and whiskers.

the first slice was that the minimum percentage of VOIs that were correctly classified was 38% instead of 0%. A similar effect could be seen in the 2nd, 3rd and 4th slice. For slice 5 and classification of the remainder of the VOIs in the scan, results were similar for both approaches. This plot indicates that using training data from previously annotated scans may only be beneficial in classification of the first axial slice. Therefore, previous training data was only used for classification of the first axial slice in protocols i-a1(-ts), i-a3(-ts), i-a3c(-ts), and i-ae(-ts).

In the bottom row of Figure 5, interactive classification results are shown for one axial slice, for observer 1 (g), observer 2 (h), and observer 3 (i). The manual VOI annotations are shown directly above. For individual observers, the classifier was trained differently, which is reflected in the results in the bottom row. This indicated that the interactive classifier was able to adapt to the annotation preferences of the observers.

3.3.2 Use of previous training data and classifier selection

In Figure 8, classification accuracy for the first slice, classification accuracy for the remainder of the scan, and overall classification accuracy are displayed, without (a) and with (b) texture selection. Median, minimum and maximum values are given in Table 6. The light grey boxes in Figure 8 represent results for classification of the remainder of the scan. Since this was done in the same way in each interactive protocol, the results do not differ between the different approaches. The results for classification of the first axial slice, displayed by the dark grey boxes, displayed a large spread of median values: from 47% for pro-

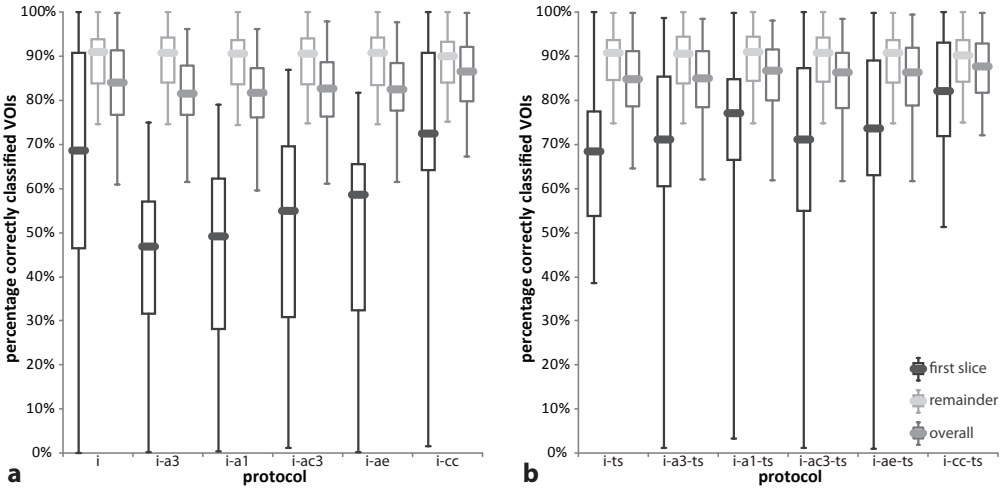


Figure 8. Box plots displaying classification accuracies of all annotated datasets for the 6 interactive protocols without texture selection (**a**) and for the 6 protocols with texture selection (**b**). For each protocol, distribution of classification accuracy of the first slice (dark grey boxes), distribution of classification accuracy of the remainder of the scan after training on 50% of all VOIs (light grey boxes), and distribution of overall classification accuracy (middle grey boxes) are shown. See the caption of Figure 6 for an explanation of the boxes and whiskers.

tocol i-a3, in which previous training data was used for classification of the first axial slice, to 82% for protocol i-cc-ts, where observers initiated annotation by selecting 5 example VOIs for each texture present in the scan and where they could choose which classification results to start from when correcting VOI labels. Protocols i-ts and i-cc-ts were the only two protocols for which the minimum number of VOIs correctly classified is not close to 0%, with 38% and 51% respectively. In the overall results, median accuracies ranged from 82% for protocol i-a3 to 88% for protocol i-cc-ts. The minimum overall percentage of correctly classified VOIs was also largest for protocol i-cc-ts: 72%. For all interactive protocols, texture selection increased overall classification accuracy by 1 to 5 percentage points.

Repeated measures ANOVA was performed to test for the difference in accuracy between the six interactive protocols without texture selection (i, i-a1, i-a3, i-a3c, i-ae, and i-cc). Protocol i-cc performed significantly better than the other ones ($F(5,28) = 29.1$; $p < 0.01$). In addition, repeated measures ANOVA was used to test the significance of texture selection on classification accuracy by comparing the accuracies of protocols i-cc and i-cc-ts. Texture selection had a significant effect on accuracy ($F(1,32) = 11.6$; $p = 0.002$). Finally, we tested whether interactive classification performed better than automatic classification by performing repeated measures ANOVA on the four automatic protocols (a1, a3, a3c, and ae) and the best-performing interactive protocol. Interactive classification performed significantly better than the four automatic protocols ($F(4,27) = 21.6$; $p < 0.001$). Post-hoc analysis indicated no significant differences between the individual automatic protocols.

Table 6. Median (min-max) % of correctly classified VOIs for all annotated datasets in the 12 interactive annotation protocols. Results are shown for classification of the first slice (top), for classification of the remainder of the scan after training on at least 50% of the VOIs in the scan (middle), and for all VOIs in the scan (bottom). a1: training data from same user; a3: training data from all users; a3c: consensus training data; ae: ensemble classification; ts: texture selection.

	Median (min-max) accuracy for the first slice					
	i	**i-a1**	**i-a3**	**i-a3c**	**i-ae**	**i-cc**
no ts	69 (0-100)	49 (0-79)	47 (0-75)	55 (1-87)	59 (0-82)	72 (2-100)
ts	68 (38-100)	77 (3-100)	71 (1-99)	71 (1-100)	74 (1-100)	82 (51-100)
	Median (min-max) accuracy for the remainder of the scan					
	i	**i-a1**	**i-a3**	**i-a3c**	**i-ae**	**i-cc**
no ts	91 (75-100)	91 (74-100)	91 (75-100)	91 (75-100)	91 (75-100)	90 (75-100)
ts	91 (75-100)	91 (75-100)	91 (75-100)	91 (75-100)	91 (75-100)	90 (75-100)
	Median (min-max) overall accuracy					
	i	**i-a1**	**i-a3**	**i-a3c**	**i-ae**	**i-cc**
no ts	84 (61-100)	82 (59-96)	82 (61-96)	83 (61-98)	83 (62-98)	87 (67-100)
ts	85 (65-100)	87 (62-98)	85 (62-98)	86 (62-98)	86 (62-99)	88 (72-100)

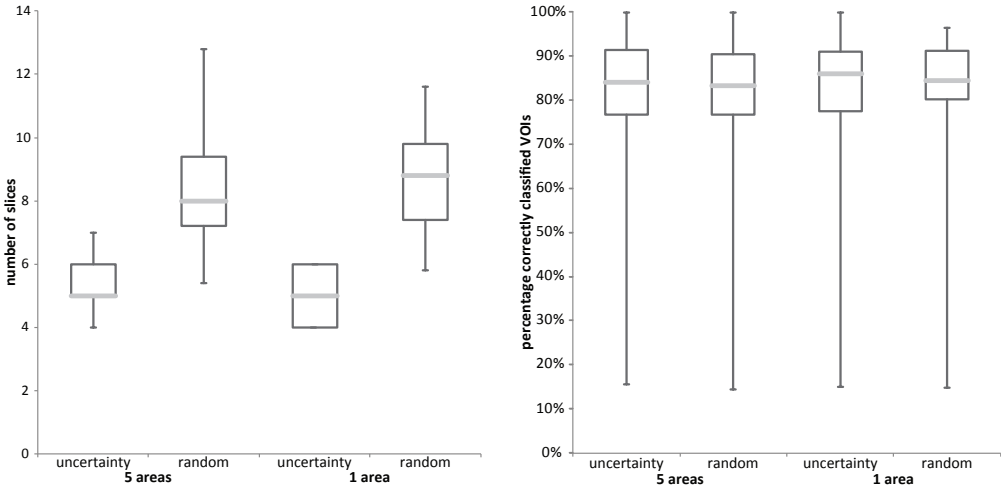


Figure 9. a. . Box plot displaying the number of slices that needed to be inspected by the observer to annotate at least 50% of all VOIs in the scan in the 4 different slice selection methods. **b.** Box plot of the overall accuracy for the 4 different slice selection methods. Experiments were performed using protocol i, without texture selection and without the use of previously annotated VOIs for classification of the first slice. See the caption of Figure 6 for an explanation of the boxes and whiskers.

3.4 Slice selection methods

Finally, we compared four different methods for slice selection in interactive classification. Figure 9a displays the distribution of the number of slices that had to be reviewed by an

observer before the threshold of 50% of the VOIs in the scan was reached. In these experiments, classification protocol i was used, in which no texture selection was performed. Both random slice selection methods required more slices to be checked; median values were 8 when the lungs were divided into 5 areas and 9 when the lungs were not subdivided. This also means that more rounds of training the classifier and reclassification of the remaining VOIs had to be done. Dividing the scan into 5 areas, from which slices were selected alternately, led to a larger spread in the number of slices that had to be reviewed. The median number of slices, 5, was the same, whether the lungs are divided into 5 areas or not. Figure 9b shows that the distribution of interactive classification accuracy per annotated dataset was comparable for all slice selection methods. Therefore, use of uncertainty-based slice selection is preferred over random slice selection. The number of areas in which the lungs were subdivided did not matter in these experiments.

4 Discussion

Many studies have focused on automatic texture analysis in ILD. In general, these studies use user-defined regions of interest (ROIs) or VOIs for which one set of annotations is used as the ground truth. In this work, we built on the interactive annotation approach that we have described earlier (Kockelkorn et al. 2016), in which all lung tissue is annotated and ground truth is defined by the observer using the software. One of our aims was to optimize the interactive annotation process.

High interobserver variability is a known issue in ILD texture annotation. In two of our previous studies, both of which were performed on smaller datasets, we found interobserver agreements for two observers of 51% (Kockelkorn et al. 2010) and 63% (Kockelkorn et al. 2016). In this work, roughly 70% of all VOIs that were labeled at least twice received the same label from at least two observers. This also means for roughly ⅓ of the VOIs, all two or three observers assigned a different label to the same VOI. With observers having three different opinions on the texture label of a VOI, the problem of making a universally accepted annotated dataset becomes even more complicated. One of the issues we wanted to study was how we could use training data from previously annotated scans for classification at the start of interactive annotation, when no or little training data from the scan under consideration is available. Given the high interobserver variability, we hypothesized that using training data from the observer that is annotating the scan might have advantages over using training data obtained from all observers. Our results indicate that using training data from the same observer results in slightly higher classification accuracy, but only when combined with texture selection. Without texture selection, using

consensus training data leads to the highest median accuracy in automatic classification. We also investigated the use of having a ensemble of classifiers deciding together on the label of VOIs, analogous to physicians making a diagnosis together. Using this ensemble approach did not lead to higher median classification accuracies when compared with the other classification protocols. In general, the differences between the median results of the various annotation methods are small and the method that yields the highest median accuracy varies for the individual observers.

Another aim of our study was to investigate how a classifier, trained on previously obtained training data, can decrease user effort necessary for interactive annotation of all lung VOIs. Since the automatic classification experiments did not indicate the superiority of one single automatic classification protocol, we tested the use of all automatic classification methods in our interactive classification framework. In the most basic protocol (i), no texture selection or automatic classification results were used. VOIs in the first axial slice were labeled as normal tissue, which is in this dataset with 55% the most common texture. The median labeling accuracy over all annotated datasets resulting from this approach was 69%, which is below median classification accuracy for all automatic protocols with class selection. Median interactive classification accuracy for all following slices is above median accuracy for all automatic protocols. Therefore, we conclude that automatic classification is only beneficial for classification of the first axial slice.

Other studies have reported higher automatic classification accuracies (Depeursinge, Van de Ville, et al. 2012, Song et al. 2013, Anthimopoulos et al. 2016), however, these were obtained on hand-drawn ROIs at specific user-selected locations instead of automatically generated VOIs spanning the entire lungs. In addition, not all classification approaches are suitable for the interactive annotation approach we propose: we opted for pre-calculated features and a classifier that is trained relatively fast to reduce the time that the observer has to wait for classification results. Nevertheless, it would be possible to initiate interactive annotation by classification of the first slice using, for example, a deep learning approach.

Besides correcting automatic classification results, observers can transfer knowledge of the annotation task in other ways. The first is by selecting examples of all tissue types present in the scan before the start of interactive classification. In automatic classification, texture selection resulted in a substantial increase of median classification accuracy, ranging from 20 percentage points for the protocol using consensus training data (a3c), to 27 percentage points for the protocol that used training data from the observer who provided the ground truth annotations for the scan under consideration (a1).

In interactive classification, using automatic classification results without texture selection leads to a decrease in classification accuracy, as compared with the protocols in which texture selection is applied. This decrease is not only seen in the first slice, but also in the overall classification results: protocols i-a1, 1-a3, i-a3c, and i-ae display a decrease in median classification accuracy of 1-2 percentage points as compared with protocol i. Therefore, automatic classification results should only be used for classification of the first axial slice if the observer performs texture selection before interactive annotation.

Finally, we noted that selecting slices to be corrected by the observer based on the cumulative uncertainty in the slice results in a smaller number of slices that should be reviewed before reaching at least 50% of the VOIs, as compared with random slice selection. A smaller number of slices to review means that the classifier has to be trained less often, which in turn decreases the time observers have to wait for new classification results to be generated. Dividing the lungs into 5 areas, from which slices are chosen in an alternate fashion, did not have advantages over considering slices from the entire lungs in each classification, retraining, and correction cycle. This was contrary to what we expected, since ILD textures may be localized.

The current work presents several open issues. First, while the 23 thoracic CT scans included are an increase as compared with our previous studies, experiments should ideally be executed on a larger dataset, containing a variety of ILD subtypes. Second, only part of the dataset was annotated by more than one observer. This enabled us to assess interobserver variability and to evaluate how the interactive annotation framework adapts to individual observers' annotation preferences. However, in order to get to a consensus dataset, three (or more) observers should individually annotate all scans and then discuss their results. Interactive annotation can facilitate this approach in future work. Third, this work does not compare the effects of using a different classifier or different features. In principle any texture features could be inserted into the interactive framework. The same holds for the classifier, with the limitation that the chosen classifier should allow training while the observer awaits the results.

In future work, it would be interesting to compare the simulation results from the current work to results obtained by human observers. Given the substantial interobserver variability, it is conceivable that observers are influenced by the classification results that are suggested by the algorithm. To investigate the influence of suggested annotations, observers could be asked to annotate the same scan twice at different time points: once by completely manual labeling of all VOIs and once by interactive labeling. By comparing the interactive annotation results to the manual labels on the one hand and the automatically

generated labels on the other hand, an estimate of the degree by which computer-generated labels influence annotation behavior of the observer could be made.

We have shown that automatic classification results can be beneficial in interactive annotation, but only when used in combination with texture selection. In addition, giving observers several different automatic classification results to choose from when correcting VOI labels decreased the median the number of corrections. Using the best-performing protocol, in which observers select the textures that should be distinguished in the scan and in which they are provided with alternative classification results in case interactive classification accuracy is low, a median accuracy of 88% was reached. We therefore conclude that interactive annotation with texture selection and classifier choice could be a useful tool for annotating lung tissue in CT scans of ILD patients.

Acknowledgments

The authors would like to thank Floor van Meer for performing the repeated measures ANOVA.

References

Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Trans Med Imaging.* **2016**;doi: 10.1109/TMI.2016.2535865.

Aziz ZA, Wells AU, Bateman ED, Copley SJ, Desai SR, Grutters JC, Milne DG, Phillips GD, Smallwood D, Wiggins J, Wilsher ML, Hansell DM. Interstitial lung disease: effects of thin-section CT on clinical decision making. *Radiology.* **2006**;238(2):725-33.

Depeursinge A, Iavindrasana J, Hidki A, Cohen G, Geissbuhler A, Platon A, Poletti PA, Müller H. Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization. *J Digit Imaging.* **2010**;23(1):18-30.

Depeursinge A, Van de Ville D, Platon A, Geissbuhler A, Poletti PA, Müller H. Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames. *IEEE Trans Inf Technol Biomed.* **2012**;16(4):665-75.

Gao M, Bagci U, Lu L, Wu A, Buty M, Shin H, Roth H, Papadakis GZ, Depeursinge A, Summers RM, Xu Z, Mollura DJ. Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization.* **2016**;doi: 10.1080/21681163.2015.1124249

Geurts P, Ernst D, Wehenkel L. Extremely Randomized Trees. *Machine Learning.* **2006**;63 (**1**):3–42

Huber MB, Nagarajan MB, Leinsinger G, Eibel R, Ray LA, Wismüller A. Huber, Markus B, Mahesh B Nagarajan, Gerda Leinsinger, Roger Eibel, Lawrence A Ray, and Axel Wismüller. Performance of topological texture features to classify fibrotic interstitial lung disease patterns. *Med Phys.* **2011**; 38(4):2035-44.

Huber MB, Bunte K, Nagarajan MB, Biehl M, Ray LA, Wismüller A. Texture feature ranking with relevance learning to classify interstitial lung disease patterns. *Artif Intell Med.* **2012**;56(2):91-7.

Kockelkorn TTJP, de Jong PA, Gietema HA, Grutters JC, Prokop M, van Ginneken B. Interactive Annotation of Textures in Thoracic CT Scans. In: *Proceedings of the SPIE.* **2010**;7624:76240X – 76240X8.

Kockelkorn TTJP, de Jong PA, Schaefer-Prokop CM, Wittenberg R, Tiehuis AM, Gietema HA, Grutters JC, Viergever MA, van Ginneken B. Semi-automatic classification of textures in thoracic CT scans. *Phys Med Biol.* **2016**;61(16):5906-24.

Park SO, Seo JB, Kim N, Park SH, Lee YK, Park BW, Sung YS, Lee Y, Lee J, Kang SH. Feasibility of automated quantification of regional disease patterns depicted on high-resolution computed tomography in patients with various diffuse lung diseases. *Korean J Radiol.* **2009**;10(5):455-63.

Prosch H, Schaefer-Prokop CM, Eisenhuber E, Kienzl D, Herold CJ. CT protocols in interstitial lung diseases--a survey among members of the European Society of Thoracic Imaging and a review of the literature. *Eur Radiol.* **2013**;23(6):1553-63.

Song Y, Cai W, Zhou Y, Feng DD. Feature-based image patch approximation for lung tissue classification. *IEEE Trans Med Imaging.* **2013**;32(4):797-808.

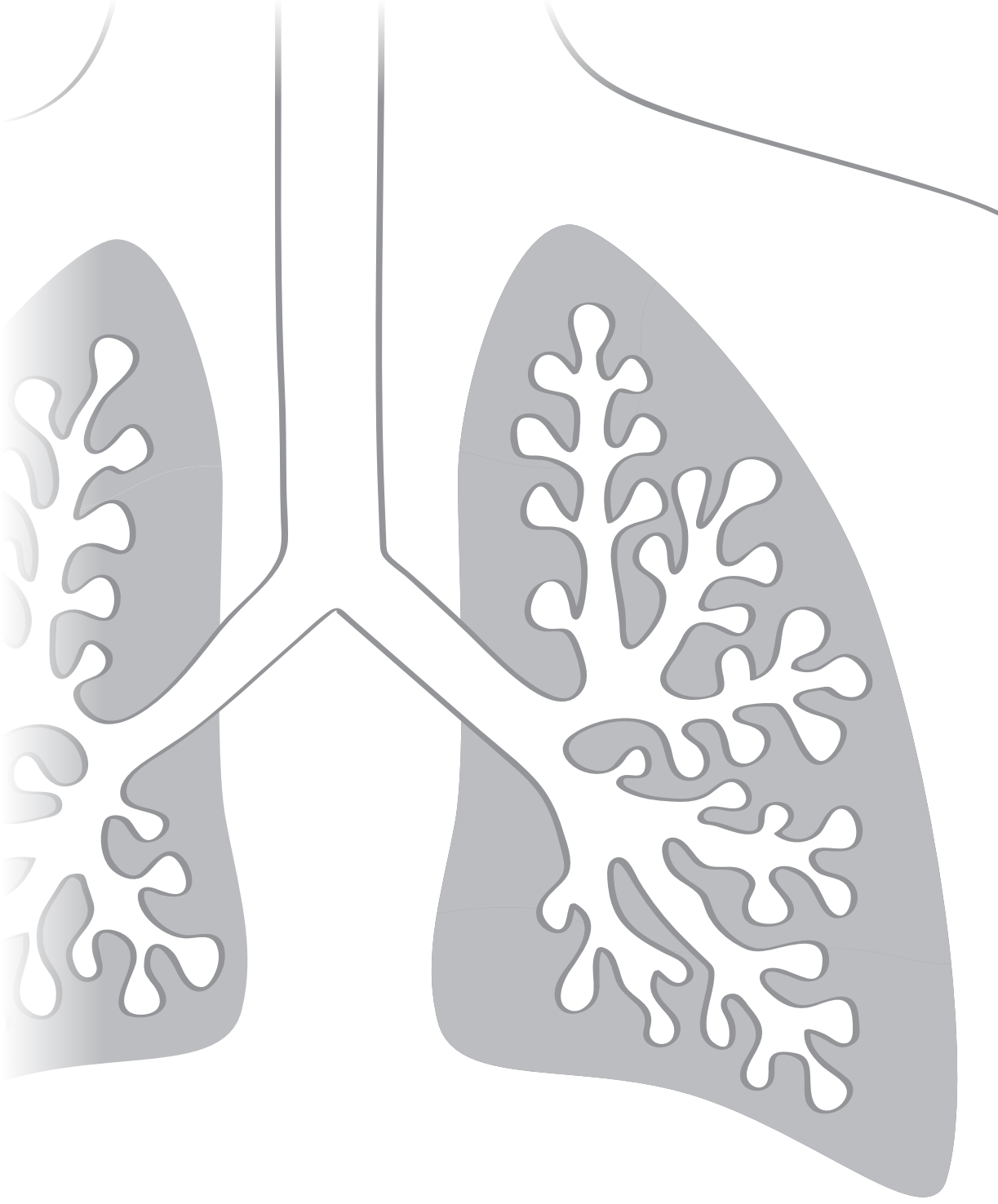
Uppaluri R, Hoffman EA, Sonka M, Hunninghake GW, McLennan G. Interstitial lung disease: A quantitative study using the adaptive multiple feature method. *Am J Respir Crit Care Med.* **1999**;159(2):519-25.

van Rikxoort EM, de Hoop B, Viergever MA, Prokop M, van Ginneken B. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Med Phys.* **2009**;36(7):2934-47.

Vasconcelos V, Barroso J, Marques L, Silva JS. Enhanced Classification of Interstitial Lung Disease Patterns in HRCT Images Using Differential Lacunarity. *Biomed Res Int.* **2015**;672520.

Xu Y, van Beek EJ, Hwanjo Y, Guo J, McLennan G, Hoffman EA. Computer-aided classification of interstitial lung diseases via MDCT: 3D adaptive multiple feature method (3D AMFM). *Acad Radiol.* **2006**;13(8):969-78.

Zavaletta VA, Bartholmai BJ, Robb RA. High resolution multidetector CT-aided tissue analysis and quantification of lung fibrosis. *Acad Radiol.* **2007**;14(7):772-87.



Interactive lung segmentation in abnormal human and animal chest CT scans

Thessa TJP Kockelkorn
Cornelia M Schaefer-Prokop
Gracijela Bozovic
Arrate Muñoz-Barrutia
Eva M van Rikxoort
Matthew S Brown
Pim A de Jong
Max A Viergever
Bram van Ginneken

Based on: Medical Physics 41, 081915 (2014)

Abstract

Purpose: Many medical image analysis systems require segmentation of the structures of interest as a first step. For scans with gross pathology, automatic segmentation methods may fail. The authors' aim is to develop a versatile, fast, and reliable interactive system to segment anatomical structures. In this study, this system was used for segmenting lungs in challenging thoracic computed tomography (CT) scans.

Methods: In volumetric thoracic CT scans, the chest is segmented and divided into 3D volumes of interest (VOIs), containing voxels with similar densities. These VOIs are automatically labeled as either lung tissue or non-lung tissue. The automatic labeling results can be corrected using an interactive or a supervised interactive approach. When using the supervised interactive system, the user is shown the classification results per slice, whereupon he/she can adjust incorrect labels. The system is retrained continuously, taking the corrections and approvals of the user into account. In this way, the system learns to make a better distinction between lung tissue and non-lung tissue. When using the interactive framework without supervised learning, the user corrects all incorrectly labeled VOIs manually. Both interactive segmentation tools were tested on 32 volumetric CT scans of pigs, mice, and humans, containing pulmonary abnormalities.

Results: On average, supervised interactive lung segmentation took under 9 min of user interaction. Algorithm computing time was 2 min on average, but can easily be reduced. On average, 2.0% of all VOIs in a scan had to be relabeled. Lung segmentation using the interactive segmentation method took on average 13 min and involved relabeling 3.0% of all VOIs on average. The resulting segmentations corresponded well to manual delineations of eight axial slices per scan, with an average Dice similarity coefficient of 0.933.

Conclusions: The authors have developed two fast and reliable methods for interactive lung segmentation in challenging chest CT images. Both systems do not require prior knowledge of the scans under consideration and work on a variety of scans.

1 Introduction

Segmentation of anatomical structures of interest is one of the first steps in most medical image analysis systems. A large body of research has focused on automated segmentation. For many organs and imaging modalities, accurate solutions have been developed. However, fully automatic segmentation of structures with prominent lesions is highly complex. We have therefore developed two flexible interactive segmentation methods for three-di-

mensional scans, that allow the user to quickly segment any type of structure, even if gross abnormalities are present. In the first method, the user corrects automatic segmentation inaccuracies one-by-one. In the second method, the system learns the characteristics of the segmentation task from the user.

In this study, we focus on lung segmentation in thoracic computed tomography (CT) scans. Existing automatic lung segmentation strategies include supervised voxel classification (Hu et al. 2001), atlas-based (Zhang et al. 2006), and model-based (Brown et al. 1997) approaches. In lungs without gross pathology, these approaches generally work well, but the presence of abnormalities typically results in segmentation errors. In 2011, the LOLA11 challenge was organized to compare state-of-the-art automatic and semi-automatic lung segmentation methods for chest CT scans on a dataset of 55 scans, that included many difficult cases (LOLA11 2011). For these difficult cases, all algorithms obtained poor results. In another recent study, Meng and co-workers collected 2768 chest CT examinations of 2292 subjects (Meng et al. 2012). These subjects suffered from a diverse set of diseases. Scans were acquired and reconstructed using different protocols and equipment. In all scans, lungs were automatically segmented using a thresholding-based approach. In 121 examinations, which corresponded to 4.4% of all cases, automatic lung segmentation performed poorly. The majority of failures were due to the presence of disease. Anatomical variations were the second largest cause of inaccuracies. Thus, these two studies indicate that there are scans in which even the best methods were not able to produce accurate lung segmentations.

One reason why automatic methods may fail to produce accurate segmentation results is the presence of large regions of hyperdense lung tissue. This tissue has attenuation values that resemble those of the tissue surrounding the lungs. Hence, segmentation schemes frequently classify these regions incorrectly as non-lung tissue. Methods that are specifically tailored to handle scans with dense lung abnormalities have been proposed (Sofka et al. 2011, Sun et al. 2012), but these algorithms still cannot make the distinction between dense lung tissue and other hyperdense abnormalities inside the thorax, such as pleural effusions and pleural thickening. In addition, these methods cannot be used out of the box on new datasets, for example, on pediatric or animal scans, without constructing a new atlas or model.

For these difficult cases, interactive methods facilitate the generation of accurate lung segmentations. Interactive segmentation methods can be based on contour-drawing or on painting (Olabarriaga and Smeulders 2001). In 2004, Kang et al. described a set of tools for interactive editing of 3D automatic segmentations in medical images (Kang et al. 2004).

They developed a hole-filling tool, a point-bridging tool, and a surface-dragging tool and used them for refinement of automatic segmentation of the proximal femur in a spiral CT dataset. Results indicated the superiority of these tools over slice-by-slice editing methods for 3D medical data. In 2010, McGuinness and O'Connor compared four interactive segmentation algorithms: seeded region growing, interactive graph cuts, simple interactive object extraction, and interactive segmentation using binary partition trees (McGuinness and O'Connor 2010). The algorithms were tested on 2D non-medical images. Twenty observers were asked to extract 100 objects from a dataset. The corresponding 100 images had been divided into 4 subsets and for each subset, observers had to use a different segmentation algorithm to segment the selected object. Their results indicated that the interactive graph cuts algorithm and the binary partition tree algorithm were most effective. For 3D medical data, such a comparison has not been made yet. Top et al. described the use of active learning for interactive segmentation of 3D images (Top et al 2011). In their method, segmentation was regarded as a classification problem. In a given segmentation, the plane where the classifier was least confident was located. This plane was shown to the user, who labeled the data on this plane and thereby provided the classifier with new training data. Using this method, user interaction time was reduced by 64% as compared to an approach in which the user chose the planes that should be labeled. El-Zehiry and co-workers proposed editing tools that use 2D contours drawn by the user for correcting 2D or 3D segmentations. These contours, together with the original image and the old segmentation were used as input for an energy minimization framework. Their system was tested on CT scans of the lungs and liver, and on MR scans of the liver. It yielded a decrease of the average segmentation error by 15% (El-Zehiry et al. 2013). Instead of editing 2D slices of automatic 3D brain segmentation results, Levinski and co-workers proposed a method for correcting errors in 3D (Levinski et al. 2009). Their method evaluated where potential errors in the initial segmentation were located and allowed users to interactively adapt the segmentation surface. Virtual reality-based frameworks have also been proposed for interactive medical image segmentation (Beichel et al. 2012, Sun et al. 2013). Results were promising, but the specialized hardware necessary for this approach is a drawback for their use in practice.

In this work, we use a different approach for interactive segmentation of the lungs. Our approach is based on predefined volumes of interest (VOIs) that are constructed to contain voxels with similar density values (Kockelkorn et al. 2010). These VOIs are automatically labeled as either lung or non-lung tissue based on their average density. There are two ways in which observers can interactively check and correct these initial automatic segmentation results. The first is by using the developed software in an interactive mode, in

which observers are asked to relabel all incorrectly labeled VOIs. The second is by using the software in a supervised interactive mode, in which observers correct errors in a slice-by-slice manner. After each slice, a classifier is iteratively trained to teach the computer the difference between lung and non-lung tissue. Slices that have not been reviewed by the observer yet are then reclassified. We compare the interactive mode and the supervised interactive mode in terms of time and user interaction required to obtain a complete lung segmentation. To show the versatility of the methods, we used scans of human subjects with gross pathology, but also scans of pigs from a lung transplantation model, and micro-CT scans of mice from a chronic pulmonary inflammation model.

2 Materials

The study comprised 32 thoracic CT scans (see Table 1) from four different origins.

The first set of scans contained eight micro-CT scans of mice with induced chronic pulmonary inflammation by silica instillation. Scans were acquired using a Micro-CAT II scanner (Siemens Pre-Clinical Solutions, Knoxville, TN), at 80 kVp X-ray source voltage and 500 μA current. Scans were reconstructed to isotropic voxels of 46 μm. The number of slices per scan varied between 466 and 570 (Artaechevarria et al. 2010).

The second set consisted of eight scans from pigs in a lung transplantation model. Scans were acquired on a 64-slice Philips Brilliance scanner at 120 kVp, with slice thicknesses ranging from 0.67–2.0 and 0.35 mm increment. In plane resolution varied between 0.34 and 0.68 mm. Images were reconstructed to 512×512 matrices or to 768×768 matrices. The number of slices was 212–1060 per scan. For the experiments described in this paper, all scans were resized to 512×512 matrices with isotropic voxels.

Table 1. Overview of the datasets used in the lung segmentation experiments. In the pig and LOLA11 scans, in plane resolutions and slice thicknesses varied among subjects. For these scans, the range is given.

Dataset name	Images from	Type of scan	# of scans	In plane resolution (mm)	Slice thickness (mm)	Range of # of slices
Mice	Murine chronic pulmonary inflammation model	Micro-CT	8	0.046	0.046	466–570
Pigs	Porcine lung transplantation model	Volumetric CT	8	0.34–0.68	0.67–2.0	212–1060
TB	Tuberculosis patients	Volumetric CT	8	0.68	1.0	280–348
LOLA11	Lung segmentation challenge	Volumetric CT	8	0.54–0.78	0.45–1.5	168–588

Eight others were scans of human tuberculosis (TB) patients, made on a Sensation Cardiac 64 scanner (Siemens Medical Solutions, Forchheim, Germany) at 120 kVp x-ray source voltage and 90 mA current. Images were reconstructed to 512×512 matrices with 1 mm slice thickness and 0.68× 0.68 mm in-plane resolution. Scans consisted of 280–348 slices.

The final eight were human scans from the LOLA11 challenge (LOLA11 2011). From the entire LOLA11 dataset, the scans containing the largest amounts of dense abnormalities were chosen. Images were acquired with several scanners using various scanning protocols, but all were reconstructed to 512×512 matrices. In-plane resolution varied between 0.54×0.54 and 0.78×0.78 mm with slice spacing between 0.45 and 1.5 mm. The number of slices per scan varied between 168 and 588. Further details of the individual scanning procedures were not available.

3 Methods

This section describes the two schemes for interactive segmentation considered in this work. A schematic overview is given in Figure 1. Both schemes rely on a subdivision of the scan under consideration into VOIs. These VOIs are automatically labeled as either lung tissue or non-lung tissue, based on their average density. In the first scheme, to which we will refer as "interactive segmentation", the user corrects the labels of all VOIs that are incorrectly labeled in the initial automatic segmentation step. In the second scheme, to which we will refer as "supervised interactive segmentation", the user is shown the automatic labeling results for one axial slice. He/she then adjusts all incorrect labels. The system is retrained continuously, taking the corrections and approvals of the user into account. In this way, the system learns to make a better distinction between lung tissue and non-lung tissue. Note that both methods are generic and could be used for many segmentation tasks, but in the description below some particularities were implemented to address the lung CT segmentation task. Subsections 3.1–3.8 will discuss the different steps of both segmentation methods and the experiments that were conducted in detail.

3.1 Thorax segmentation

For all scans, the thorax was automatically segmented with a simple algorithm. First, the field of view (FOV) was determined. Starting from the corners of the image, the percentage of voxels with the same value was calculated for each line of voxels in the z-direction. If the percentage was smaller than 90%, the line of voxels is considered to be inside the FOV. If this percentage was at least 90%, the line of voxels is ruled to be outside the FOV and its eight-connected neighbors were inspected. Second, the chest was segmented by

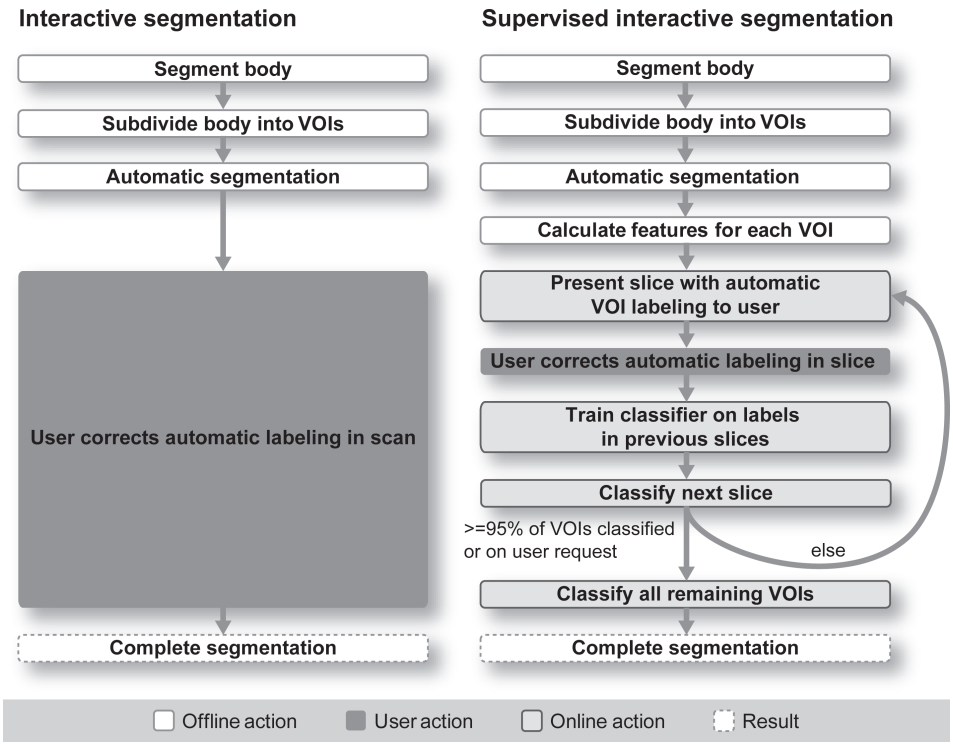


Figure 1. Flowcharts of the processes of interactive segmentation left and supervised interactive segmentation (right).

taking each axial slice and thresholding it at -200 Hounsfield units (HU). In the mask containing all voxels above this threshold, connected component labeling using four-connected neighbors was performed and the largest component was kept. This corresponds roughly to the tissues outside the lungs. Hole filling was applied to this largest component to find all voxels belonging to the chest, including the lungs.

3.2 VOI calculation

Each chest segmentation was subdivided into roughly spherical VOIs, containing voxels with similar density values, according to a scheme that we have described previously (Kockelkorn et al. 2010). To this end, the original scans were downsized to 256×256 matrices with isotropic voxels, and afterwards blurred using a Gaussian kernel with $\sigma = 1$ voxel. In the down-sampled scans, local minima and maxima within the chest segmentation were selected as seeds for growing the VOIs. A minimum distance between the seeds, indicated in Table 2, was chosen to limit the number of VOIs that were formed. Since the datasets contained scans from different subject populations that were acquired with different scan-

ners and protocols, an appropriate minimum distance was chosen for each dataset. The distance parameter represents a trade-off between segmentation effort and segmentation accuracy. A lower distance leads to more and smaller VOIs and therefore to more user effort in the segmentation task. A higher value leads to larger VOIs, and hence a higher probability of more than one texture appearing in one VOI. This results in less precise segmentations. To each VOI seed, all voxels inside the chest segmentation within a radius of ⅓× the minimum seed distance were added to form the initial VOI. For all six-connected neighboring voxels of all initial VOIs, a dissimilarity score was calculated, indicating the difference between the voxel and its neighboring VOI. This dissimilarity score was based on the difference between the voxel density value and the average density of the initial VOI, and on the distance from the voxel to the seed of the VOI as given by

$$D = \left|(H_v - \overline{H})\right| + C \times d^2 \qquad (1)$$

where *Hv* denotes the density of the voxel in HU, $\overline{H}$ is the average density value in HU in the initial VOI, and *d* is the distance in mm from the voxel to the seed of the VOI. The constant *C* denotes the relative weight assigned to the absolute difference in distance. A smaller *C* puts more emphasis on the difference in density, which leads to a more homogeneous density distribution and a more irregular shape of the VOI. A larger *C* results in compact VOIs with more spread in density values. In this study, we chose values for *C* which yielded roughly spherical VOIs, as assessed by visual inspection. The chosen values for this relative weight *C* are given in Table 2. All voxels neighboring the growing VOIs were kept in a sorted list *L*. Since all VOIs were grown simultaneously, this list contained all unassigned voxels neighboring at least one growing VOI. The construction of VOIs was therefore a competitive process. We assumed that the voxel with the lowest dissimilarity score was most likely to belong to its adjacent VOI. Therefore, it was added to the VOI and removed from *L*. For all voxels neighboring this newly added voxel which had not been assigned to a VOI already, and for which the voxel/VOI combinations were not contained in *L*, dissimilarity scores were calculated. These neighboring voxels were added to *L*. Again, the voxel with the lowest dissimilarity score was added to its adjacent VOI and new voxels were added to *L* if necessary. This process was repeated until all voxels in the chest segmentation were assigned to a VOI. The image containing the VOIs was resized to the original scan size and the borders of the VOIs were slightly smoothed afterwards. This was done by taking the most common label in a 3×3×3 neighborhood for each voxel in the resized image. Please note that in Eq. (1), only differences in intensity are taken into account, whereas a texture criterion is absent. However, results from our previous study indicate that this current method yields VOIs with homogeneous textures (Kockelkorn et al. 2010).

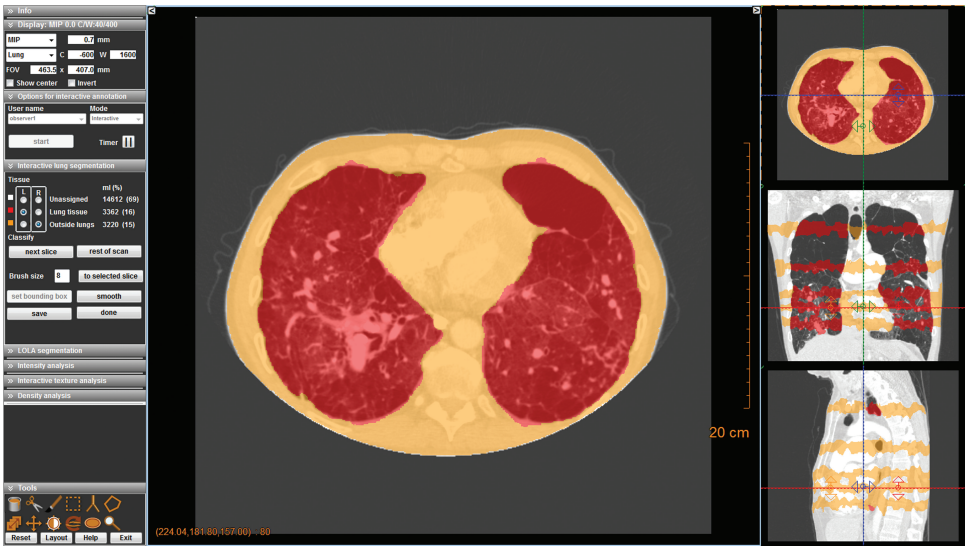


Figure 2. Screenshot of the segmentation environment in supervised interactive mode. Several slices have been classified by the system and corrected by the user. The axial, coronal, and sagittal images on the right hand side show the 3D orientation of the slice in the main window.

3.3 Automatic labeling

In both segmentation protocols, VOIs were initially labeled as lung tissue or other tissue based on their average density in the original scan. VOIs with an average density below or equal to −500 HU were labeled as lung tissue, the rest as outside the lungs. This provided a reasonable starting point for both interactive lung segmentation protocols.

3.4 Features and classifier

In the supervised interactive protocol, a *k*-nearest neighbor (*k*-NN) classifier (Duda et al 2001) with $k = 7$ was retrained iteratively to learn the characteristics of the segmentation task from the user. For each VOI, 72 features were calculated. Each scan was filtered using Gaussian, Laplacian, and gradient magnitude filters and the three eigenvalues of the Hessian at three scales ($\sigma = 1$, 2, and 4 voxels), yielding 18 filtered images. For each VOI in each filtered image, the average, standard deviation, skew, and kurtosis of the voxel values were calculated. Note that these features are rotationally invariant, as the textures that need to be classified do not have a specific orientation. Features were normalized before classification.

3.5 Supervised interactive segmentation

In the supervised interactive protocol, the user was shown an axial slice with VOI labels, as determined by the initial VOI labeling step. Labels were shown as a semitransparent overlay, with different colors for lung and non-lung tissue. As can be seen in Figure 2, labels were shown for all voxels in all VOIs intersecting with this slice. The user was asked to inspect 3D VOIs spanning several axial slices at once, which made this approach faster than the examination of each individual axial slice. VOIs in the remainder of the scan were not covered by the overlay and were referred to as "unassigned". The user corrected all errors made in the labeled slice by using the computer mouse. To each mouse button, one tissue class (lung or non-lung) could be assigned. By clicking on a voxel in the axial slice, the underlying VOI was assigned the label corresponding to the mouse button that was used. Upon dragging, all encountered VOIs received the label that had been assigned to the pressed mouse button. If the user was satisfied with the segmentation, all VOIs intersecting with the slice under consideration were assumed to have a correct label.

All assigned VOIs were used as training data for reclassification of the unassigned VOIs. After classification, a second axial slice was chosen. Again, the user corrected all errors. The VOIs intersecting with this slice were added to the training data. The classifier was retrained and all unassigned VOIs were reclassified. Slices were chosen in such an order to sample the scan at different levels, as can be seen in the coronal and sagittal views in Figure 2. To this end, the scan was divided into nine parts, each containing an equal number of axial slices in the z-direction. Each time, a different part of the scan was visited and neighboring parts were not visited consecutively. From the selected part, the axial slice with the largest number of unassigned voxels was chosen for classification.

The process of classification and correction continued until the labels of at least 95% of all voxels were approved by the observer or until the observer terminated the slice-by-slice process, typically because he/she believed the results were almost correct. The remainder of the scan was then classified and the completely segmented scan was shown. The observer could scroll through the scan in any direction, make the final corrections, and save the resulting lung segmentation.

3.6 Interactive segmentation

In the non-supervised interactive protocol, the user was shown the entire scan with automatically generated labels for all VOIs. He/she had to correct all errors in the entire scan, by clicking and dragging in the same fashion as in the interactive protocol.

3.7 Reference standard

To evaluate the accuracy of the segmentation results, we used manual delineations of eight axial slices of each scan (Hu et al. 2001, Korfiatis et al. 2008, Wang et al. 2009, van Rikxoort et al. 2009). Four of these slices were selected by the first author. In these slices, lung segmentation was challenging, typically because of the presence of dense or other abnormalities. The other four were chosen at 1st, 2nd, 3rd, and 4th of the total height of the lungs, in order to sample them at different levels. Ground truth segmentations for the LOLA11, TB, and pig scans were made by the first author and checked by a thoracic radiologist. For the mice scans, reference lung segmentations were provided by Artaechevarria et al. (2009). These segmentations were made using the interactive tools of AMIRA (Stalling et al. 2005). In the slices selected for evaluation, the interactive segmentations were further refined by the first author and checked by a thoracic radiologist.

3.8 Experiments

For the experiments, we used the segmentation environment shown in Figure 2. In all scans, lungs were segmented by observer A and observer B independently. Both observers were medical students, who were trained for this task. Each observer segmented each scan twice in two different sessions: once using the interactive segmentation mode and once using the supervised interactive segmentation mode. For half of the scans, users were instructed to first follow the interactive protocol, for the other half, they first had to follow the supervised interactive protocol. The order in which scans were segmented was random and different for both sessions. Interactive and supervised interactive segmentation of the same scan took place on different days. As can be seen in Table 3, for observer A on average 5 days passed between the segmentation of the same scan with the different protocols. For observer B, this was 4 days on average.

The primary outcome measure of this research is the user effort required for lung segmentation. User effort was measured in two ways: by the time it took to get to a complete lung segmentation and by the number of times that a user changed a VOI label during segmentation. Both metrics were recorded automatically by the software. In addition, we assessed segmentation accuracy, interobserver agreement, and intraobserver agreement. To evaluate accuracy, we compared all segmentations to manual delineations of eight axial slices in each scan. Interobserver agreement was estimated by comparing the segmentation results of the two observers in these same eight slices of each scan, both for the interactive and the supervised interactive protocol. Intraobserver agreement was estimated by comparing the interactive and the supervised interactive segmentation results of each

Table 2. Overview of the parameters for VOI creation for each dataset.

Dataset name	Minimum seed distance (mm)	Relative weight (C in Eq. (1))	Number of VOIs (average (range))
Mice	0.75	100	9516 (7654–11227)
Pigs	4	20	13690 (9361–18328)
TB	6	3	10417 (7417–12106)
LOLA11	10	3	12815 (8687–18362)

Table 3. Average number of days elapsed between two segmentations of the same scan by observer A and B.

Dataset name	Observer A	Observer B
Mice	5	3
Pigs	2	3
TB	5	5
LOLA11	8	4
Total	5	4

observer, again in these eight reference slices. As an evaluation measure, we used the Dice similarity coefficient (DSC),

$$DSC = 2\frac{|A \cap B|}{|A| + |B|} \tag{2}$$

where A denotes the set of voxels in the first segmentation and B the set of voxels in the second segmentation.

4 Results

4.1 Segmentation effort

The average number of VOIs and their ranges per dataset are given in Table 2. In Figure 3(a) and Table 4, the number of label changes by the user are given for both segmentation protocols. Values are averaged over both observers and over all scans per dataset. Interactive segmentation required on average more label changes than supervised interactive segmentation, for all datasets. Standard deviations, as indicated by the whiskers, are high, since the number of VOIs needing relabeling varied per scan within a dataset. They are lower for the supervised interactive than for the interactive protocol. Table 4 gives a more detailed overview of the number of label changes necessary for a complete segmentation. On average, interactive segmentation required 332 label changes, which corresponds to relabeling 3.0% of all VOIs. Scans from the LOLA11 challenge required the smallest per-

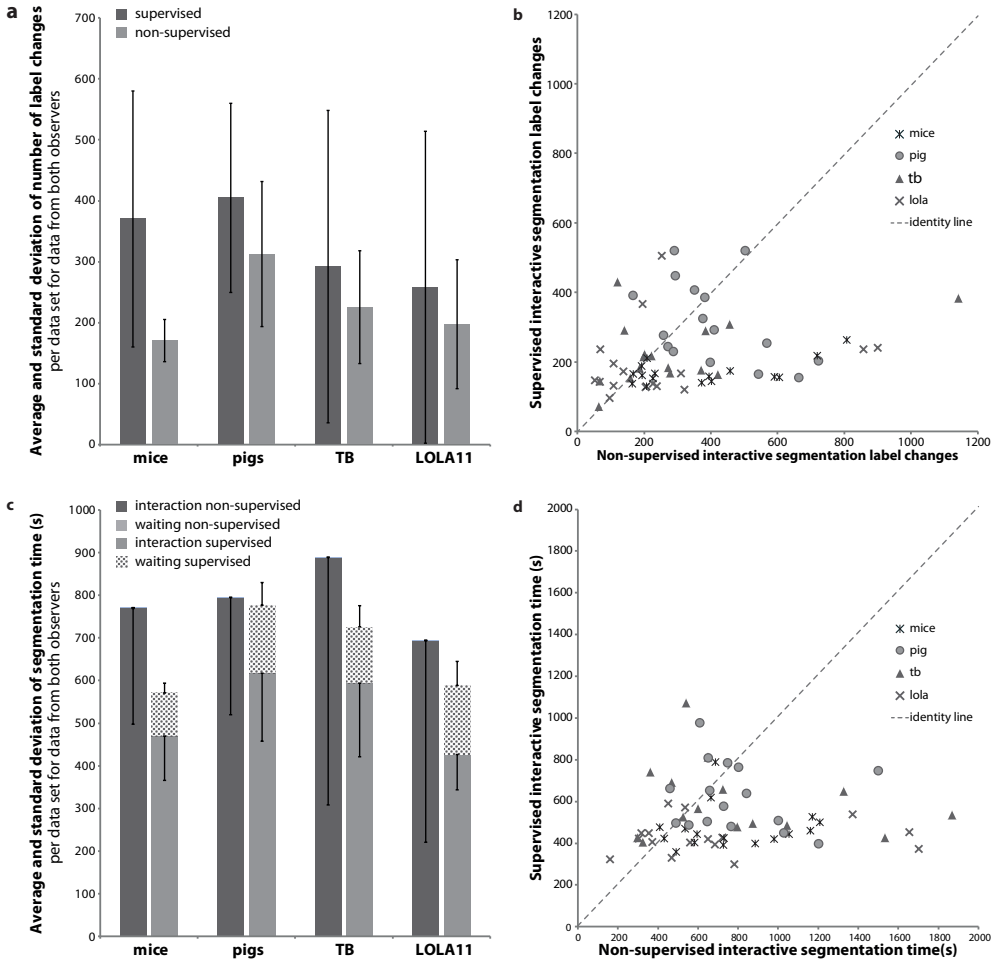


Figure 3. (Left) bar charts of the number of label changes per dataset (**a**) and of the segmentation time in seconds (**c**) in the interactive protocol and the supervised interactive protocol. In each bar, values are averaged over both observers and over all cases per dataset. The whiskers indicate the standard deviations. In (**c**), separate bars display the total segmentation times and the user interaction times for the supervised interactive protocol. (Right) scatter plots of the number of label changes (**b**) and of the user time (**d**) required for interactive (horizontal) versus supervised interactive (vertical) segmentation. Each data point represents one scan segmented by one observer.

centage of changes: 2.2% or 259 VOIs and the mouse scans needed the most changes: 4.0%, corresponding to 371 VOIs. Segmentation of the two most normal scans in both the LOLA11 and TB database required on average relabeling of 0.5% of the VOIs. Supervised interactive segmentation took on average 227 label changes, which corresponds to relabeling 2.0% of the VOIs in a scan. This is a decrease of labeling effort of 32% as compared with the interactive segmentation protocol. The benefits of the supervised interactive system

are largest in the mouse scans: the supervised interactive protocol required on average 171 label changes, which is a decrease of 54% as compared with the interactive protocol. This difference is significant ($p < 0.05$; paired t-test), as is the difference between the number of label changes in the interactive and the supervised interactive protocol for all datasets taken together.

Figure 3(b) shows a scatter plot indicating the number of label changes necessary to get to a complete lung segmentation in the supervised interactive (vertical) and the interactive (horizontal) protocol. Each data point corresponds to one scan, segmented by one observer. Most data points are located below the diagonal identity line, which means that for these cases, supervised interactive segmentation required fewer label changes than interactive segmentation. For the cases which are above the line, the opposite holds. Here, manual correction of all VOI labels required less user interaction than using an iteratively retrained classifier. For scans that require higher numbers of label changes when using the interactive method, supervised interactive segmentation offered a larger reduction in label changes. The cases in which the use of the supervised interactive system caused an increase in the number of label changes were also the cases which required a lower number of manual corrections. In other words, these are the cases in which the automatic labeling was already quite accurate. Thus the supervised interactive system offered most benefit for the more difficult cases.

Abnormalities in the lungs are expected to be an important reason for the initial threshold-based labeling method to fail. We calculated the percentage mislabeled lung VOIs by counting the number of VOIs incorrectly labeled as non-lung in the initial labeling step and dividing it by the number of VOIs categorized as lung in the ground truth segmentation. The results, averaged over all scans per dataset, both observers and both segmentation protocols, are displayed in Table 5. Most errors were made in the initial segmentations of mouse and pig scans, where over 10% of all VOIs categorized as lung in the reference segmentations were incorrectly labeled as non-lung. The initial segmentation method worked better for the TB and LOLA11 scans. In contrast, VOIs that were located outside the lungs according to the ground truth segmentation were less likely to receive an incorrect label from the initial segmentation method.

Figure 3(c) shows the time in seconds needed to segment the lungs using the interactive and the supervised interactive protocol. Times presented are averaged over both observers and all scans per dataset. For the supervised interactive protocol, the total time (including waiting time), and user interaction time are displayed in separate bars. On average, interactive segmentation took longer than supervised interactive segmentation. Standard

Table 4. Label changes per dataset, averaged over all cases in a dataset and over both observers. The percentage label changes in the columns labeled "Interactive" and "Supervised interactive" are the absolute numbers of label changes divided by the average number of VOIs per scan, multiplied by 100. The absolute gain is calculated as the number of label changes in the interactive protocol minus those in the supervised interactive protocol. The percentage gain is the absolute gain divided by the number of changes in the interactive protocol, multiplied by 100. SD = standard deviation; abs = absolute; * indicates that this difference is significant ($p < 0.05$; paired t-test).

Dataset name	Average number of VOIs	Interactive		Supervised interactive		Gain	
		abs (SD)	% (SD)	abs (SD)	% (SD)	abs	%
Mice	9516	371 (211)	4.0 (2.2)	171 (35)	1.8% (0.3)	200	* 54
Pigs	13690	406 (156)	3.2(1.6)	313 (119)	2.3% (0.8)	93	23
TB	10417	292 (257)	2.7(2.1)	226 (93)	2.2% (0.9)	67	23
LOLA11	12815	259 (257)	2.2(2.4)	197 (106)	1.6% (0.6)	61	24
Total	11609	332 (226)	3.0 (2.1)	227 (106)	2.0% (0.7)	105	* 32

Table 5. Percentages of VOIs that were incorrectly labeled as non-lung by the initial threshold-based labeling step per dataset. Results were averaged over all scans per dataset, both observers and both segmentation protocols.

Dataset name	Percentage (%)
Mice	12
Pigs	11
TB	4
LOLA11	5
Total	7

Table 6. Segmentation times per dataset (in seconds), averaged over all cases in a dataset and over both observers. Total segmentation time can be split into user time and waiting time. User time is defined as the time that the user interacted with the segmentation software. Waiting time is the time during which the user had to wait for processing steps of the system to be finished. In the interactive segmentation protocol, waiting time was negligible and was therefore left out of the table. The column titled "User time gain abs" displays the difference in user time for the interactive and the supervised interactive segmentation protocol. The column titled "Total time gain abs" displays the difference in user time for the interactive and total time for the supervised interactive segmentation protocol. The percentage time gain is the absolute time gain divided by the interactive segmentation time. SD = standard deviation; abs = absolute; * indicates that this gain is significant ($p < 0.05$; paired t-test) for both observers.

Dataset name	Interactive	Supervised interactive		User time gain		Total time gain	
	User time (SD)	User time (SD)	Total time (SD)	abs	%	abs	%
Mice	770 (272)	469 (104)	568 (115)	300	* 39	202	26
Pigs	794 (275)	617 (160)	775 (184)	177	22	21	2
TB	889 (581)	593 (173)	724 (214)	295	33	166	19
LOLA11	693 (473)	426 (83)	588 (94)	267	39	105	15
Total	786 (417)	526 (155)	664 (179)	260	* 33	124	**16**

Table 7. Dice similarity coefficients indicating the agreement between the segmentations made by the observers and the reference standard per dataset, per observer, and per segmentation protocol. Results are averaged over all eight slices per scan for which manual delineations were made and over all scans in the dataset.

	Observer A		Observer B	
Dataset Name	Interactive	Supervised interactive	Interactive	Supervised interactive
Mice	0.937	0.933	0.939	0.938
Pigs	0.904	0.902	0.909	0.908
TB	0.932	0.937	0.952	0.950
LOLA11	0.948	0.947	0.948	0.947
Total	0.930	0.929	0.937	0.936

deviations, indicated by whiskers, are lower for the supervised interactive than for the interactive protocol. Table 6 shows the times needed for segmentation in more detail. Interactive segmentation took on average 786 s, with a minimum of 693 s for LOLA11 scans and a maximum of 889 s for the TB scans. The time during which the user had to interact with the supervised system was 526 s, with a range from 426 s for the LOLA11 scans to 617 s for the pig scans. Time gained by using the supervised interactive system ranged from 22% for the pig scans to 39% for the mouse model and LOLA11 datasets. The time gain for the mouse scans was significant for both observers, as was the time gain averaged over all datasets. In the interactive segmentation protocol, waiting time was negligible. In the supervised interactive protocol, the user had to wait on average 137 s. This waiting time could easily be reduced by optimizing the software, for example, by using a multi-threaded implementation. If waiting time is also included in the analysis, a time gain of on average 16% was achieved.

Figure 3(d) shows a scatter plot of supervised interactive segmentation time, without waiting time, (vertical) versus interactive (horizontal) segmentation time. Similar to Figure 3(b), supervised interactive segmentation required less time than interactive segmentation. For scans which took long in the interactive segmentation protocol, supervised interactive segmentation offered a larger reduction in segmentation time.

4.2 Segmentation accuracy

Table 7 shows the DSC calculated for the segmentations made by the two observers and the manual delineations of eight difficult axial slices in each scan. Results were averaged over all reference slices in all scans in each dataset. The bottom row displays the average DSC over all datasets. Overall, the DSC ranged between 0.902 and 0.952, which indicates a good correspondence between the manual delineations and the segmentations that were

Table 8. Dice similarity coefficients indicating interobserver and intraobserver agreement, calculated over the eight reference slices. DSC was averaged over all scans in a dataset. The row labeled "total" displays the DSC averaged over all datasets. Interobserver agreement was estimated by calculating the DSC of the segmentations of observer A and observer B, both for the interactive protocol (column labeled "A versus B interactive") and for the supervised interactive protocol (column labeled "A versus B supervised interactive"). Intraobserver agreement was assessed by calculating the DSC of the resulting segmentations of the interactive and those of the supervised interactive segmentation method, for observer A ("interactive versus supervised interactive A") and for observer B ("interactive versus supervised interactive B").

	Interobserver agreement A versus B		Intraobserver agreement Interactive versus supervised interactive	
Dataset Name	Interactive	Supervised interactive	A	B
Mice	0.979	0.982	0.979	0.985
Pigs	0.970	0.978	0.974	0.978
TB	0.972	0.978	0.985	0.984
LOLA11	0.992	0.993	0.993	0.995
Total	0.978	0.983	0.983	0.986

made using our interactive software. DSCs were comparable for both protocols. The difference in DSC between both segmentation protocols for the TB dataset as segmented by observer A was statistically significant ($p < 0.05$; paired t-test). All other differences were not statistically significant.

In Table 8, interobserver and intraobserver agreements are given. Interobserver agreement was assessed by calculating the DSC of the results of observer A and observer B when using the interactive protocol (in the column named "A versus B interactive") and using the supervised interactive protocol (in the column labeled "A versus B supervised interactive"). The DSCs were calculated over the same eight slices per scan that were also used to evaluate the accuracy of segmentation. The DSC averaged over all datasets was comparable for both protocols, with values of 0.978 for the interactive and 0.983 for the supervised interactive approach. The range in DSCs for the individual datasets is small.

Intraobserver agreement was evaluated by calculating the DSC of the segmentations made using the interactive and of those made using the supervised interactive segmentation method, both for observer A (in the column labeled "interactive versus supervised interactive A") and for observer B (in the column labeled "interactive versus supervised interactive B"). The average DSC over all datasets was 0.983 for observer A and 0.986 for observer B. The spread in DSCs for the different datasets is small.

5 Discussion

We have presented two interactive segmentation frameworks, one supervised and the other non-supervised, which can be used for a variety of complex segmentation tasks. We have demonstrated the usefulness of both frameworks to segment lungs in four datasets for which an established automatic solution was not available. This is the case for scans containing large abnormalities or for those of animal models. Building an automatic system is labor-intensive and as long as the number of scans to be processed is relatively low, using a versatile interactive segmentation system is an attractive alternative.

With the supervised interactive segmentation protocol we presented, lungs could be segmented in 11 min on average, by relabeling 2.0% of all VOIs. Of these 11 min, users needed to wait 2 min for the computer to finish calculations. When using a multithreaded implementation, this waiting time can be reduced by a factor equal to the number of processing cores. In the interactive protocol, correcting all errors in the automatic segmentation manually took 13 min on average, for relabeling 3.0% of all VOIs. For comparison: manual delineation of one lung fiels in one axial slice takes approximately 1 min. This means that for a typical scan, in which around 400 slices contain lung tissue, 13 h of manual segmentation would be needed (Sluimer et al. 2005). This process can be sped up by using interpolation methods to reduce the number of slices that are delineated. However, we could not estimate from literature how much time can be saved in this way. In addition, the use of interpolation methods is usually at the expense of accuracy.

We compared the segmentation results to manual delineations of eight slices in each scan. Example segmentations for each dataset, as made by both observers using both protocols, are displayed in Figure 4. The lung segmentations obtained by both methods showed a good correspondence to these manual delineations with a DSC varying between 0.902 and 0.952. Intraobserver and interobserver correspondence as estimated by the DSC were higher, on average 0.982. The higher DSCs in Table 8 (interobserver and intraobserver agreement) as compared to Table 7 (segmentation accuracy) may have two causes: lung borders are not followed accurately by the VOI borders or both observers did agree with themselves and with each other, but not with the reference standard. Abnormal dense areas around the lung borders may be caused by a variety of processes, both inside and outside the lungs, which makes interpretation of these areas nontrivial. These subjective judgments can cause large interobserver variability, but are unrelated to the choice for a particular segmentation tool. Figure 4 shows an axial slice of one scan per dataset. Row (b) displays the manual delineations of the lungs, which were checked by a radiologist. As can be seen in rows (c), (d), (e), and (f) for the LOLA11 slice, which are the segmentation

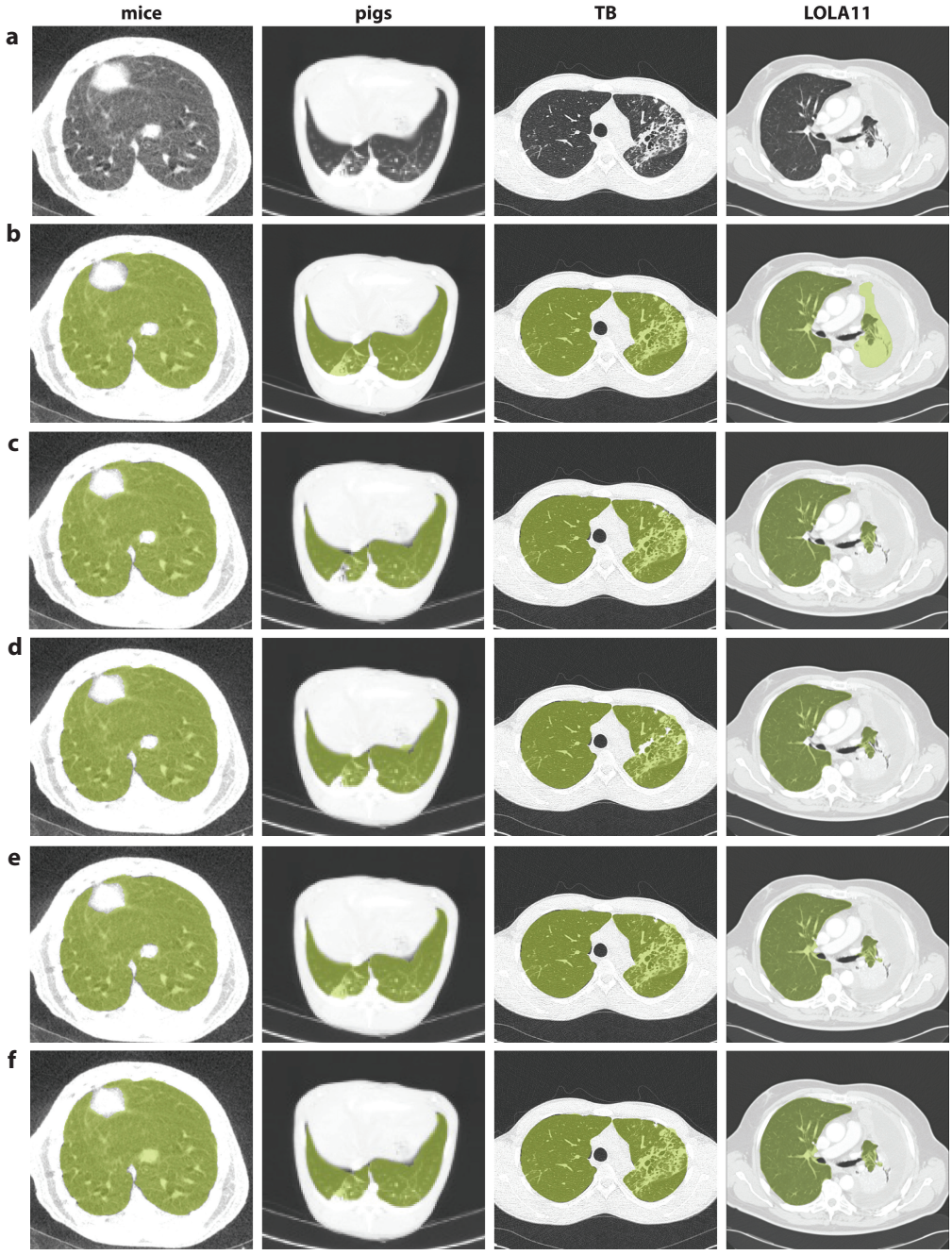


Figure 4. Examples of the different segmentation results in an axial slice from one CT scan per dataset. Segmentations are displayed as semitransparent overlays. Row (**a**) axial slice without any overlays. Row (**b**) manual delineations of both lungs as verified by a radiologist. Row (**c**) segmentation by observer A using the supervised interactive segmentation approach. Row (**d**) segmentation by observer A using the interactive segmentation approach. Row (**e**) segmentation by observer B using the supervised interactive segmentation pproach. Row (**f**) segmentation by observer B using the interactive segmentation approach.

results of the observers, both did not agree with the radiologist on the exact location of the lung boundaries. We therefore conclude that at least part of the difference between the DSCs for agreement with the manual delineations and the DSCs indicating interobserver and intraobserver agreement are due to the fact that the observers in our experiments did not agree with the radiologist on the exact location of the lung borders. Similarly, the example segmentations of the axial pig slice show that accurate segmentation of abnormalities is possible with the proposed segmentation method [in row (e)], but that the final segmentation result is guided by the subjective criteria of the observer. Therefore, we conclude that the subdivision of a scan in VOIs is accurate, in the sense that the resulting VOIs follow the lung borders.

These results indicate that our system offers a fast and accurate way to perform lung segmentations in the absence of a satisfying automatic solution. Especially for difficult cases, the supervised interactive system offers large benefits. In easier cases, manually correcting the errors in the automatic segmentation is faster and requires less user interaction. In the supervised interactive protocol, the initial automatic segmentation is only used for labeling of the first axial slice that is shown to the user. In the following slices, classification results of a *k*-NN classifier are shown. This classifier needs to be trained and becomes more accurate when more training samples are added, i.e., when the user has inspected more axial slices. As a result, in the first slices, the classifier is less accurate than the automatic segmentation method. We think that in more difficult scans, this initial disadvantage of the *k*-NN classification method is compensated by better results in later slices. In easier cases, this disadvantage of having to train the *k*-NN classifier has a negative effect on the final number of labels that have to be changed. When using the interactive approach, the observer is not limited to the slice-by-slice processing that is used in the supervised interactive setup, but he/she can freely scroll through the scan in all directions. The advantages of both systems could be combined in a system in which users could first view the automatically obtained segmentation results and then decide on a per case basis whether they want to use the supervised interactive or the interactive approach. This choice was not offered in the experiments described in this paper, but it may further reduce the user effort needed to segment a set of scans.

The described methods yield adequate segmentations for all four datasets, but the degree to which the supervised interactive method required less user effort as compared with the interactive method differed. The difference in label changes between both modes was largest for the mouse scans. For the other three datasets, the decrease in label changes between the interactive and the supervised interactive protocol was smaller. This can be explained by the fact that the mouse scans were made using a micro-CT scanner. They contain more

noise and therefore, the initial automatic labeling was less accurate. Here, using the supervised interactive mode caused a large decrease in both time and label changes the user had to invest to segment the lungs.

The described framework is highly versatile, offering possibilities for segmentation of other structures as well as possibilities for incorporating techniques that further improve segmentation ease and accuracy. Any structure which can be subdivided into roughly spherical VOIs would be amenable to segmentation by our approach. If a structure is thin and elongated, adaptations to the algorithm for VOI creation might be made by changing the value of *C* in Eq. (1). Other tasks for which this interactive system can be used are segmentation of the liver or brain structures in humans. The fact that the system can already be used for mouse and pig scans indicates that segmentation of structures in other animals is also feasible. The approach we describe can also be used for other tasks than segmentation, such as interactive texture analysis in diseased lungs (Kockelkorn et al. 2010). Future work will focus on this application.

The framework as presented in the current manuscript suffers from a number of limitations. We propose three possible adaptations to further improve segmentation accuracy and user-friendliness.

First, segmentation ease and accuracy can be improved by adapting the procedure by which the VOIs are constructed. In previous experiments, we have used the algorithm for VOI construction to distinguish different types of abnormal textures in the lungs of patients with interstitial lung disease. Two radiologists independently used the software to annotate all VOIs within the lungs. To assess the homogeneity of the VOIs, both were also asked to label VOIs that contained more than one type of texture. One observer labeled 1.2% of all VOIs as heterogeneous, the other none. Nevertheless, the method we used for calculating the VOIs is not always able to exactly follow the borders of the structures that need to be segmented.

The first possible problem is the creation of the initial VOIs by adding all voxels in a sphere with a certain diameter from the VOI seed. If the seed is close to or even on an edge of a structure, the initial VOI will automatically contain two types of tissue and the resulting segmentation will not be precise at this region. This can be counteracted by constructing smaller VOIs or by allowing users to split VOIs if necessary, but both solutions would lead to more user interaction. A second problem arises when growing the VOIs by adding neighboring voxels. One term in Eq. (1) is the distance from a neighboring voxel to the seed of the VOI. The constant *C* in front of this term can be optimized to force VOIs to be more or less spherical. This makes the shape of the VOIs more predictable for the user,

but it is also more likely that tissue borders are not exactly followed. The other element that is currently included in Eq. (1) is density. In certain applications, texture might be an additional criterion for calculating the dissimilarity between a voxel and its neighboring VOI. Since texture cannot be calculated for one voxel, a workaround is needed, for example calculating how texture features of the growing VOI would change if that specific voxel was added. The downside of this approach is that it would involve more complicated calculations, leading to increased computation time.

Second, the proposed interactive segmentation frameworks can be improved by training them on the results of previous segmentation sessions. In contrast with most interactive segmentation methods, one of our segmentation methods is supervised. During a segmentation session, it progressively learns how to segment the structure of interest from the user. In practice, the user often needs to segment a number of scans for a particular experiment. Results from previous interactive segmentation sessions can be used as training data for the initial automatic VOI-based segmentation, replacing the rather crude procedure based on thresholding that was used in this study. Increasing amounts of training data will lead to a better initial segmentation and therefore reduce the amount of user interaction necessary.

Third, improvement of the classification results could be achieved by changing the classifier and feature set used in the supervised interactive segmentation method. In this paper, a simple classifier was used. More advanced classifiers, feature sets and feature selection methods can easily be plugged into this framework and lead to an improved system that needs less training data to perform well.

6 Conclusions

We have presented two versatile interactive systems for segmenting lungs in human and animal scans, for which a proven automatic segmentation method does not exist. The presented systems offer a fast way to accurately segment lungs. These segmentation methods can easily be adapted for other medical segmentation tasks.

References

Artaechevarria X, Pérez-Martín D, Reinhardt JM, Ortiz-de-Solorzano C. Automated quantitative analysis of a mouse model of chronic pulmonary inflammation using micro x-ray computed tomography. In: *Proceedings of the Second International Workshop on Pulmonary Image Analysis*, London, UK. **2009**.

Artaechevarria X1, Blanco D, Pérez-Martín D, de Biurrun G, Montuenga LM, de Torres JP, Zulueta JJ, Bastarrika G, Muñoz-Barrutia A, Ortiz-de-Solorzano C. Longitudinal study of a mouse model of chronic pulmonary inflammation using breath hold gated micro-CT. *Eur Radiol.* **2010**;20(11):2600-8.

Beichel R, Bornik A, Bauer C, Sorantin E. Liver segmentation in contrast enhanced CT data using graph cuts and interactive 3D segmentation refinement methods. *Med Phys.* **2012** Mar;39(3):1361-73.

Brown MS, McNitt-Gray MF, Mankovich NJ, Goldin JG, Hiller J, Wilson LS, Aberle DR. Method for segmenting chest CT image data using an anatomical model: preliminary results. *IEEE Trans Med Imaging.* **1997**;16(6):828-39.

Duda RO, Hart PE, Stork DG, *Pattern Classification*, 2nd ed. (John Wiley and Sons, New York, **2001**).

El-Zehiry N, Jolly M.-P, Sofka M. A splice-guided data driven interactive editing. In: *IEEE International Symposium on Biomedical Imaging.* **2013**:1098–1101.

Hu S, Hoffman EA, Reinhardt JM. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE Trans Med Imaging.* **2001**;20(6):490-8.

Kang Y, Engelke K, Kalender WA. Interactive 3D editing tools for image segmentation. *Med Image Anal.* **2004**;8(1):35-46.

Kockelkorn TTJP, de Jong PA, Gietema HA, Grutters JC, Prokop M, van Ginneken B, Interactive annotation of textures in thoracic CT scans. In: *Proc. SPIE .* **2010;7624:**76240X-1–76240X-8.

Korfiatis P, Kalogeropoulou C, Karahaliou A, Kazantzi A, Skiadopoulos S, Costaridou L. Texture classification-based segmentation of lung affected by interstitial pneumonia in high-resolution CT. *Med Phys.* **2008**;35(12):5290-302.

LOLA 11 challenge **2011**, URL: http://lola11.com. Last accessed on July 6, 2012.

Levinski K, Sourin A, Zagorodnov V. Interactive surface-guided segmentation of brain MRI data. *Comput Biol Med.* **2009**;39(12):1153-60.

McGuinness K, O'Connor NE. A comparative evaluation of interactive segmentation algorithms. *Pattern Recog.* **2010;43**:434–444

Meng X, Qiang Y, Zhu S, Fuhrman C, Siegfried JM, Pu J. Illustration of the obstacles in computerized lung segmentation using examples. *Med Phys.* **2012**;39(8):4984-91. doi: 10.1118/1.4737023.

Olabarriaga SD, Smeulders AW. Interaction in the segmentation of medical images: a survey. *Med Image Anal.* **2001**;5(2):127-42.

van Rikxoort EM, de Hoop B, Viergever MA, Prokop M, van Ginneken B. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Med Phys.* **2009** Jul;36(7):2934-47.

Sluimer IC, Prokop M, van Ginneken B. Toward automated segmentation of the pathological lung in CT. *IEEE Trans Med Imaging.* **2005**;24(8):1025-38.

Sofka M, Wetzl J, Birkbeck N, Zhang J, Kohlberger T, Kaftan J, Declerck J, Zhou SK. Multi-stage learning for robust lung segmentation in challenging CT volumes. *Med Image Comput Comput Assist Interv.* **2011**;14(Pt 3):667-74.

Stalling D, Westerhoff M, Hege H-C. Amira: A highly interactive system for visual data analysis. **2005**.

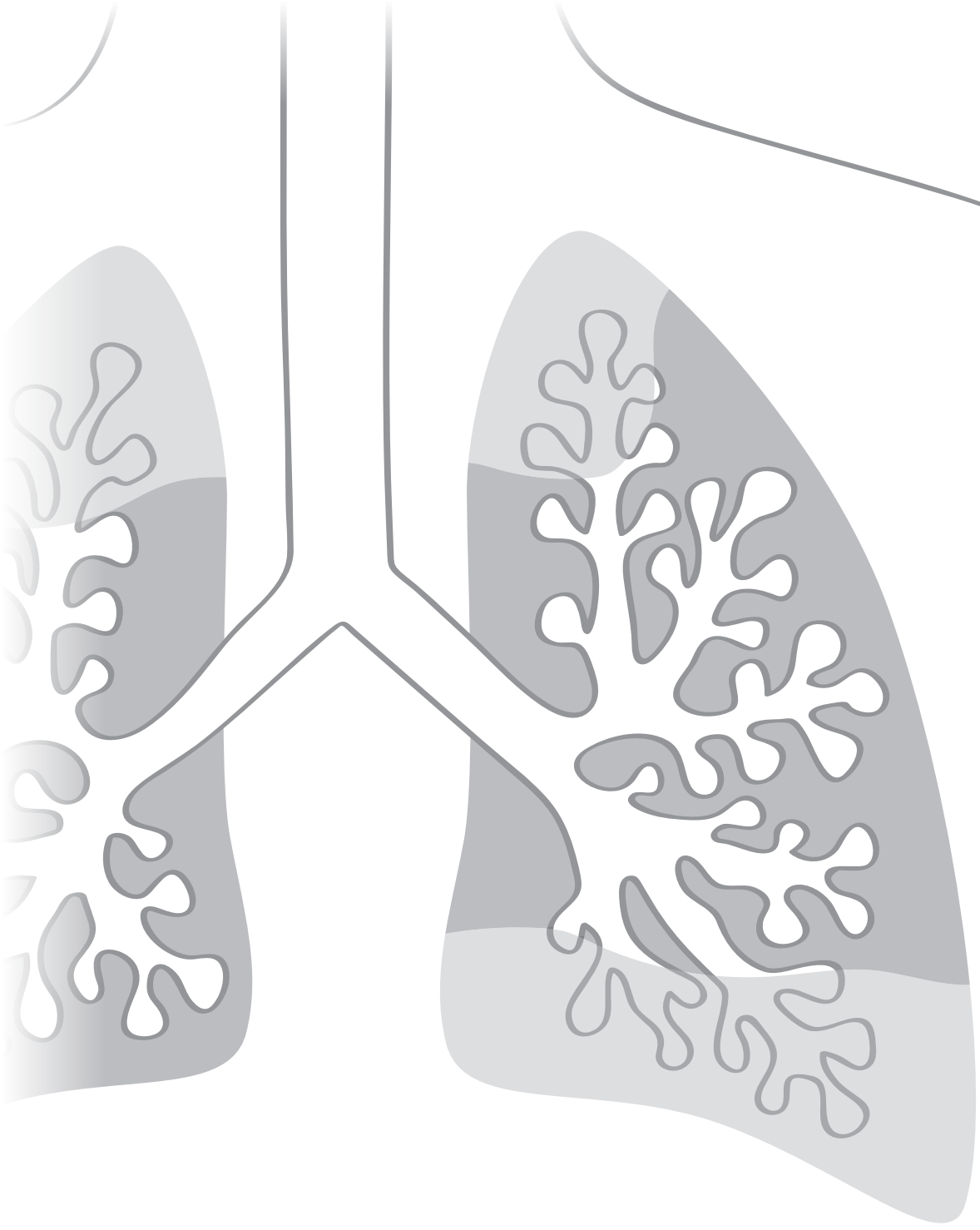
Sun S, Bauer C, Beichel R. Automated 3-D segmentation of lungs with lung cancer in CT data using a novel robust active shape model approach. *IEEE Trans Med Imaging.* **2012**;31(2):449-60.

Sun S, Sonka M, Beichel RR. Lung segmentation refinement based on optimal surface finding utilizing a hybrid desktop/virtual reality user interface. *Comput Med Imaging Graph.* **2013**;37(1):15-27.

Top A, Hamarneh G, Abugharbieh R. Active learning for interactive 3D image segmentation. *Med Image Comput Comput Assist Interv.* **2011**;14(Pt 3):603-10.

Wang J, Li F, Li Q. Automated segmentation of lungs with severe interstitial lung disease in CT. *Med Phys.* **2009**;36(10):4592-9.

Zhang L, Hoffman EA, Reinhardt JM. Atlas-driven lung lobe segmentation in volumetric X-ray CT images. *IEEE Trans Med Imaging.* **2006**;25(1):1-16.



Interactive measurement of aerated lung volume in CT scans of ICU patients

Thessa TJP Kockelkorn
Dave A Dongelmans
Ludo F Beenen
Cornelia M Schaefer-Prokop
Pim A de Jong
Max A Viergever
Bram van Ginneken

In preparation

Abstract

Background Although mechanical ventilation is indicated when spontaneous breathing is inadequate, it puts patients at risk for developing further lung injury. Lowering the applied tidal volume decreases that risk. In current clinical practice, the only factors taken into account when calculating tidal volume are patient height and presence of acute respiratory distress syndrome. The lung volume that is still able to perform gas exchange is likely to be a better predictor for determining optimal tidal volume. In this proof-of-principle study, we developed an interactive method for determining total lung capacity (TLC) and for classification of lung tissue into four aeration categories: hyperinflated, normally aerated, poorly aerated, and non-aerated lung tissue.

Methods Thirty chest CT scans of patients admitted to the intensive care unit (ICU) were collected retrospectively and lungs were segmented interactively. Total lung capacity was calculated as the gas content of all voxels included in the lung segmentation. Interactively measured TLC was compared with TLC estimated based on patient height. Lungs were subdivided into smaller volumes of interest (VOIs) with similar density values. All VOIs were automatically assigned to one of the aeration categories based on their median CT number. To test the reliability of automatic categorization, an observer was asked to refine category assignments interactively. Volumes of the four aeration categories were calculated as the total gas content of all voxels per category. Interactively obtained hyperinflated, normally aerated, poorly aerated, and non-aerated lung volume were used as a reference standard and were compared with automatic voxel-wise volume estimations for the aeration categories, and with automatic VOI-based estimations. Finally, the correlation between applied tidal volume and tidal volume calculated based on patient height was calculated, as well as the correlation between applied tidal volume and interactively calculated normally aerated lung volume, and the correlation between applied tidal volume and the interactively calculated sum of normally and poorly aerated lung volume.

Results In this group of patients, TLC calculated based on patient height was overestimated by 3.7 liter on average when compared with CT-based measurement of TLC. Results for individual patients varied widely, from underestimation of 0.4 l to overestimation of 6.6l. Interactive measurement of normally aerated lung volume deviated even more from TLC predicted based on patient heights. For some patients, interactive refinement of automatic VOI categorization was necessary to obtain a reliable calculation of normally and poorly aerated lung volume. No correlation was found between the applied tidal volume and the interactively measured normally aerated lung volume.

Discussion TLC, hyperinflated, normally aerated, poorly aerated, and non-aerated lung volume could be determined interactively. Neither TLC nor normally aerated lung volume could be estimated reliably based on body height in our patient group The proposed interactive quantification method may serve as a tool for further investigation of the relation between normally aerated lung volume and optimal tidal volume in mechanical ventilation.

1 Introduction

Mechanical ventilation is indicated when a patient's own respiration is inadequate. This inadequacy may have a wide variety of reasons, for example chronic obstructive pulmonary disease (COPD), atelectasis, neurological diseases, neurological damage, cancer, extensive pneumonia, alveolar damage, or acute respiratory distress syndrome (ARDS). ARDS in itself is not a disease, but a manifestation of underlying pathology, which is characterized by inflammation throughout the lungs. This inflammation results in injury to cells in the alveolar barrier, surfactant dysfunction, and abnormal coagulation. As a result, gas exchange in the lungs is impaired at alveolar level (Fanelli and Ranieri 2015). While mechanical ventilation is a potentially life-saving intervention, it may lead to added damage to the lungs via so-called ventilator-associated lung injury (VALI). Several mechanisms are thought to be responsible for VALI: overdistension of alveoli as a result of ventilation with high volumes, and atelectrauma, which occurs when lower volumes are used and alveoli are alternately opened (recruited) and closed (derecruited) (Slutsky and Ranieri 2015). A recent concept that is closely related to the use of tidal volumes is driving pressure. This concept leans on the idea that the amount of energy that is applied to the lung causes damage. Using this concept leads to individualized ventilator setting (Amato et al. 2015). However, this emerging concept is not the current paradigm and even the current paradigm of protective ventilation is not adopted everywhere. In most ICU patients, parts of the lung are not aerated due to pre-existing conditions. This varies from atelectasis to for instance pneumonia. The remaining aerated lung tissue is more likely to be overdistended and thus damaged, even if protective ventilator settings are applied.

Figure 1 presents a visual overview of the scope of this paper. The top part of the figure is a graph of the lung volumes that are considered. Tidal volume (V_T) is the volume of air moved into or out of the lungs during normal breathing. Total lung capacity (TLC) is the volume of air in the lungs at maximal inspiration. In order to decrease the probability that patients with ARDS develop VALI, ventilation strategies using a lower tidal volume are used (middle part of figure 1). These tidal volumes are calculated based on patient body

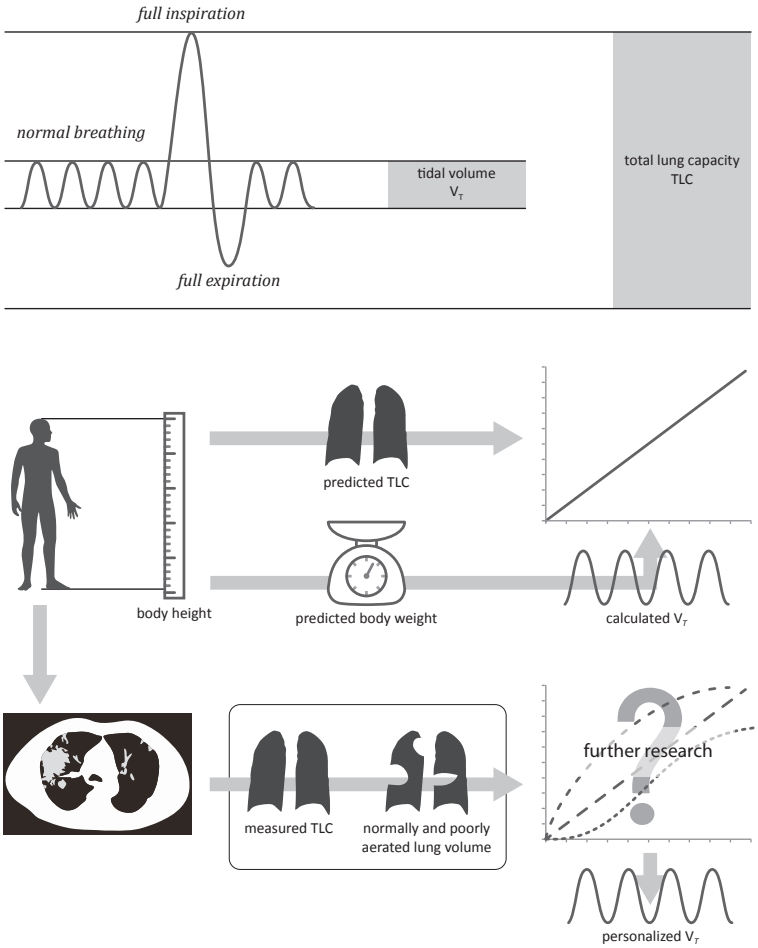


Figure 1. Top: graph displaying tidal volume and total lung capacity. Middle: Current methods to estimate tidal volume for ICU patients. Bottom: proposed method for calculation of personalized tidal volume. The rounded rectangle with black border in the bottom of the figure contains the steps covered in this paper.

height. From body height, the predicted body weight (PBW) is deduced (Devine 1974). For most patients, initially a tidal volume of 8 ml per kg of predicted body weight (PBW) is used (American Thoracic Society et al. 1999). For patients with ARDS, the maximum recommended tidal volume is 6 ml per kg of PBW. These values have been shown to result in reduced mortality in patients with ARDS (The Acute Respiratory Distress Syndrome Network 2000). Based on the available studies, no conclusions on morbidity could be drawn. However, the recommended multiplication factors do not take the extent of lung injury into account, and safe tidal volumes cannot be determined for all patients using only height as an input parameter. The same tidal volume in a patient with a massive pneumonia, resulting in 50% non-aerated lung tissue, will have a different effect in another patient

who, due to atelectasis, has 90% aerated lung tissue left. In the latter case, the applied tidal volume maybe truly protective, whilst in the former, tidal volume should be further reduced. Therefore the spatial and volumetric distribution of areas of various lung densities is of large importance for optimizing ventilation parameters. Such information is impossible to gain from bedside radiographs and requires cross-sectional imaging such as CT.

1.1 Related work

In the past, various research groups have set out to optimize ventilation settings in order to prevent VALI. Quantitative analysis of the lungs in CT scans may provide means to diagnose ARDS and optimize mechanical ventilation settings (Puybasset et al. 2000, Gattinoni et al. 2001, Malbouisson et al. 2001, Rouby et al. 2003, Schreiter et al. 2004, Rylander et al. 2004, Patroniti et al. 2005, Gattinoni et al. 2006, Caironi et al. 2010). The general approach is to classify lung voxels in one of the following categories, based on their CT value in Hounsfield units (HU): hyperinflated, normally aerated, poorly aerated, and non-aerated. The exact limits used to distinguish the different classes vary between the different studies. Reske and coworkers proposed an amendment of the approach proposed by Gattinoni (2001). Instead of evaluating the entire scan, Reske et al. used only 10 CT sections for quantification. Results were extrapolated to the entire scan. They reported a good agreement between results derived from the 10 selected slices and results from the entire scan. In this way, both radiation exposure and time required for making manual segmentations could be reduced (Reske et al. 2010).

1.2 Contributions

Despite the possible advantages of these quantification techniques, they have not been widely adopted in clinical practice. This hesitance may in part be attributed to the possible disadvantages, which include patient discomfort, risks of transportation to the CT scanner, costs, radiation exposure, and amount of time that clinicians have to invest to manually segment the lungs in the CT scans in order to facilitate quantitative analysis (Gattinoni and Cressoni 2010). To overcome this last problem, we propose to use an interactive method for analysis of chest CT scans of ICU patients, to provide personalized mechanical ventilation settings in the future (bottom part of Figure 1). By choosing an interactive approach, the time experts have to invest to obtain measurements is decreased as compared with a manual approach, while quality of the measurements is ensured. The method consists of two steps. First, lungs in the CT scans are interactively segmented. From this step, TLC can be derived. Second, lung segmentations are further annotated using the following labels: hyperinflated, normally aerated, poorly aerated, or non-aerated. From this step, the normally and poorly aerated lung volume can be derived.

Table1. Patient and scan characteristics.

Scan nr	Patient sex	Patient age (y)	In-plane resolution (mm)	Slice spacing (mm)	Tube current (mA)	Scan nr	Patient sex	Patient age (y)	In-plane resolution (mm)	Slice spacing (mm)	Tube current (mA)
1	f	64	0.736	1.00	120	16	f	54	0.586	1.00	120
2	m	65	0.705	0.75	120	17	m	43	0.551	0.75	100
3	m	72	0.848	0.75	120	18	m	79	0.586	3.00	120
4	f	61	0.693	3.00	100	19	f	77	0.611	0.75	120
5	f	44	0.781	0.75	100	20	f	19	0.625	0.75	120
6	m	80	0.785	3.00	120	21	f	34	0.600	0.75	120
7	m	65	0.756	0.75	100	22	m	57	0.631	0.75	120
8	m	57	0.664	1.00	100	23	m	70	0.730	0.75	100
9	m	57	0.742	3.00	100	24	m	52	0.621	0.75	100
10	m	63	0.600	0.75	100	25	m	77	0.719	0.75	120
11	f	79	0.586	0.75	120	26	m	58	0.828	0.75	100
12	m	81	0.713	0.75	100	27	m	63	0.729	1.30	120
13	m	50	0.631	0.75	120	28	m	63	0.682	1.00	120
14	m	45	0.734	3.00	120	29	m	57	0.768	3.00	120
15	m	79	0.826	0.75	100	30	m	40	0.586	1.00	120

Section 2 describes the scans that were used. Section 3 gives an overview of the approach we used to determine the degree of aeration of the lungs. Section 4 lists the main results, which are further discussed in section 5. Finally, in section 6 we draw conclusions from this study.

2 Materials

In total 44 standard chest CT scans of consecutive patients admitted to the intensive care unit (ICU) at the Academic Medical Center (Amsterdam, the Netherlands) between October 2008 and March 2010 were collected retrospectively. Of these, 30 were suitable for further analysis. Scans were excluded if patients were not mechanically ventilated or if clinical data was missing. Scan and patient characteristics are summarized in Table 1. All scans were acquired on a Siemens Sensation 64 scanner (Siemens Healthcare, Forchheim, Germany), except for scan 28, which was acquired on a 4-slice Philips Mx8000 scanner (Philips Healthcare, Eindhoven, the Netherlands). Fifteen patients suffered from pneumonia (patients 1-15 in table 1); the others did not (patients 16-30 in table 1).

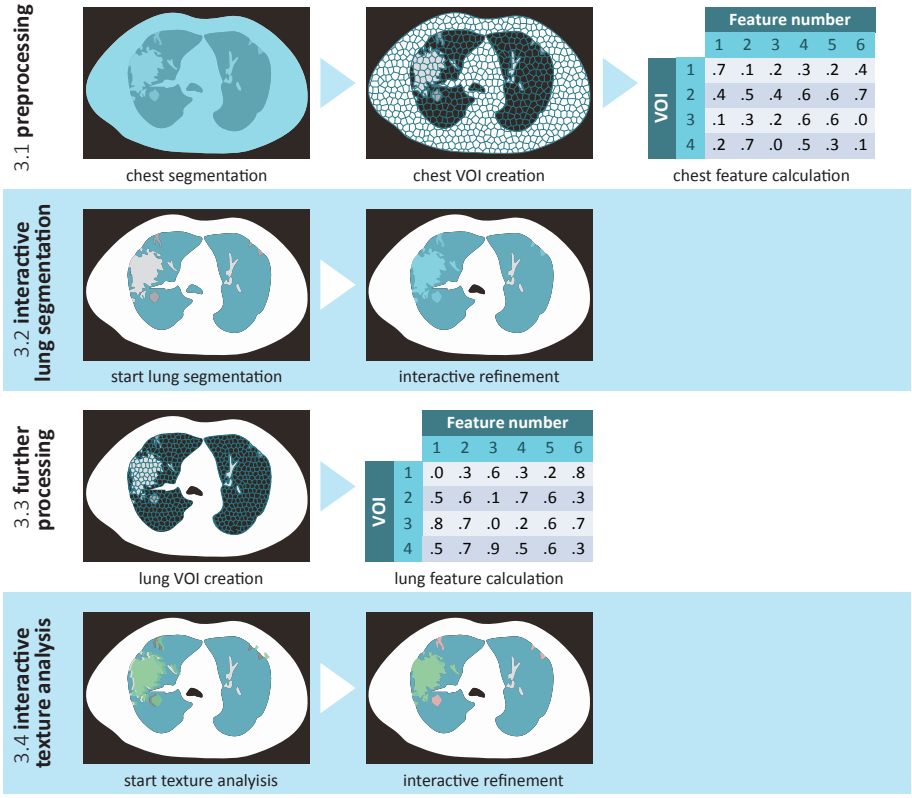


Figure 2. Pipeline of interactive aeration analysis. White rectangles represent automatic preprocessing steps. Blue rectangles represent interactive analysis.

3 Methods

Figure 2 gives a schematic overview of the different steps of the algorithm we developed to determine the aerated lung volume. The two main steps are lung segmentation and interactive texture analysis of the lungs. Each of these steps is preceded by preprocessing. In the following paragraphs, all elements of the pipeline are described in detail.

3.1 Preprocessing

Scans 3, 7 and 21 were manually cropped before preprocessing, since they included body parts that were not relevant for this study. The chest was then automatically segmented and divided into volumes of interest containing homogeneous texture following a previously proposed procedure (Kockelkorn et al. 2014). Scans were downsized to 256×256 matrices with isotropic voxels, and afterwards blurred using a Gaussian kernel with $\sigma = 1$ voxel. In the downsized and blurred scans, local minima and maxima which were at least

12 mm apart were selected as VOI seeds. All voxels within a sphere of 4 voxels were added to these seeds to form the initial VOIs. VOIs were further grown in a competitive fashion. All voxels neighboring a VOI received a dissimilarity score *D*, which was calculated as follows:

$$D = \left|(H_v - \overline{H})\right| + C \times d^2 \qquad (1)$$

where H_v was the CT value of the voxel in HU, $\overline{H}$ denoted the average CT value in HU of the initial VOI, and d was the distance in mm between the voxel and the seed of its neighboring VOI. The voxel with the lowest dissimilarity score was added to its neighboring VOI. Dissimilarity scores were calculates for the new neighboring voxels which had not yet been assigned to a VOI, and for which the voxel/VOI combinations were not contained in *L*. Again, the voxel with the lowest score was added to its neighboring VOI and new dissimilarity scores were calculated. This process continued until all voxels in the chest were included in a VOI.

3.2 Interactive lung segmentation

Since the lungs of the patients included in this study in general contained a considerable amount of various forms of pathology, an automatic lung segmentation method would probably fail in many cases. Therefore, lungs were segmented interactively (Kockelkorn et al. 2014). To this end, all VOIs were automatically labeled as either lung tissue or non-lung tissue, based on their average CT number. All VOIs with an average density equal to or below -500 HU were assumed to contain lung tissue; all other VOIs were labeled as tissue outside the lungs. These labels were presented to a medical student, with an option to refine them in two ways: 1) by changing the labels of all incorrectly labeled VOIs one by one, or 2) by iteratively training a classifier to distinguish lung tissue from other tissue by correcting automatic classification results in a slice-by-slice manner. The first method was used for scans for which the automatic labeling method yielded a reasonable lung segmentation. The observer manually corrected the labels of all incorrectly labeled VOIs. If automatic labeling was inaccurate, the second method was used as a starting point for segmentation of one axial slice. The observer corrected all mislabeled VOIs in this slice and all VOIs in the slice were used as training data for a classifier. That classifier classified the labels of the VOIs intersecting with a second axial slice, after which the classifier was retrained using all VOIs in the slices already presented to the observer. This process of classification, correction, and retraining continued until the observer decided that the classifier was fully trained and he prompted classification of all remaining VOIs in the scan. Alternatively, if at least 95% of all VOIs had been reviewed in this way, all remaining VOIs were classified automatically. In both cases, the observer reviewed the scan one final

time before saving the segmentation results. All segmentations were later reviewed by an experienced chest radiologist and corrected if necessary.

3.3 Interactive aeration analysis

The lung segmentations obtained in the previous step were subdivided into smaller VOIs according to the procedure described in (Kockelkorn et al. 2016). The smaller VOIs were automatically labeled based on their median CT value in HU as follows:

hyperinflated from -1000 to -900 HU
normally aerated from -900 to -500 HU
poorly aerated from -500 to -100 HU
non-aerated from-100 to 100 HU

In similarity with the process of interactive lung segmentation, these initial aeration labels were interactively refined by an experienced chest radiologist. He corrected the labels of incorrectly labeled VOIs.

TLC was measured in the CT scans of all patients by automatical counting of the number of voxels in the corresponding lung segmentations and multiplying voxel counts by voxel volumes and voxel gas content ($TLC_{measured}$). Gas content (GC) was calculated using the following formula (Gattinoni et al. 2005):

$$GC = -\frac{CTnumber}{1000} \qquad (2)$$

where CT number is in HU. CT numbers > 0 were set to 0.

Volumes of hyperinflated, normally aerated, poorly aerated, and non-aerated lung tissue were calculated in three ways: 1) by assigning all lung voxels to one of the categories based on their CT number (voxel-based$_{automatic}$), 2) by assigning the voxels in each VOI to one of the categories based on the median CT number within the VOI (VOI-based$_{automatic}$), and 3) by counting all voxels per aeration category as obtained by interactive labeling (VOI-based$_{interactive}$). In all cases, the resulting voxel counts were multiplied by voxel volumes and gas content.

3.4 Analysis of clinical data

For all patients, height was known. PBW in kg was calculated according the following formulas:

$$PBW = 45.5 + 0.91 \times (height - 152.4) \qquad (3)$$

for females and

$$PBW = 50 + 0.91 \times (height - 152.4) \quad (4)$$

for males. Height is expressed in cm. The tidal volume used in mechanical ventilation was calculated by multiplying PBW by 6 ml per kg PBW for patients with and 8 ml per kg PBW for patients without ARDS (American Thoracic Society et al. 1999, The Acute Respiratory Distress Syndrome Network. 2000).

Total lung capacity (TLC) was estimated from patient height using the following formulas:

$$TLC_{predicted} = 6.60 \times height - 5.79 \quad (5)$$

for females and

$$TLC_{predicted} = 7.99 \times height - 7.08 \quad (6)$$

for males. Height is expressed in cm (Quanjer et al. 1993).

3.5 Experiments and evaluation

$TLC_{predicted}$ was compared to $TLC_{measured}$, to the normally aerated lung volume as determined by the VOI-based$_{interactive}$ method, and to the sum of the normally and poorly aerated lung volume as determined by the VOI-based$_{interactive}$ method. Results were visualized in Bland-Altman plots. Normally aerated lung volume, as calculated by the voxel-based$_{automatic}$ and the VOI-based$_{automatic}$ techniques, were compared to normally aerated lung volume, as determined by the VOI-based$_{interactive}$ method, in Bland-Altman plots. Finally, the correlation between applied tidal volume on the one hand, and calculated tidal volume, normally aerated lung volume, and the sum of normally and poorly aerated lung volume as determined by the VOI-based$_{interactive}$ method on the other hand was evaluated using scatter plots.

4 Results

Figure 3a shows a Bland-Altman plot of $TLC_{predicted}$ based on patient heights versus TLC as measured in CT scans. On average, determining $TLC_{predicted}$ led to an overestimation of 3.7 l as compared with $TLC_{measured}$. This average was the same for patients with and without pneumonia. The spread in differences for individual patients was large, ranging from an underestimation of 0.4 l to an overestimation of 6.6l.

In patients with pre-existing lung injury, normally aerated lung volume is more important than TLC. After lung segmentation, all lung voxels were classified into the four aeration categories by using the voxel-based$_{automatic}$, the VOI-based$_{automatic}$ and the VOI-based$_{interactive}$ methods. In figure 3b, predicted TLC was compared with normally aerated lung volume as measured using the VOI-based$_{interactive}$ method. The deviations between predicted

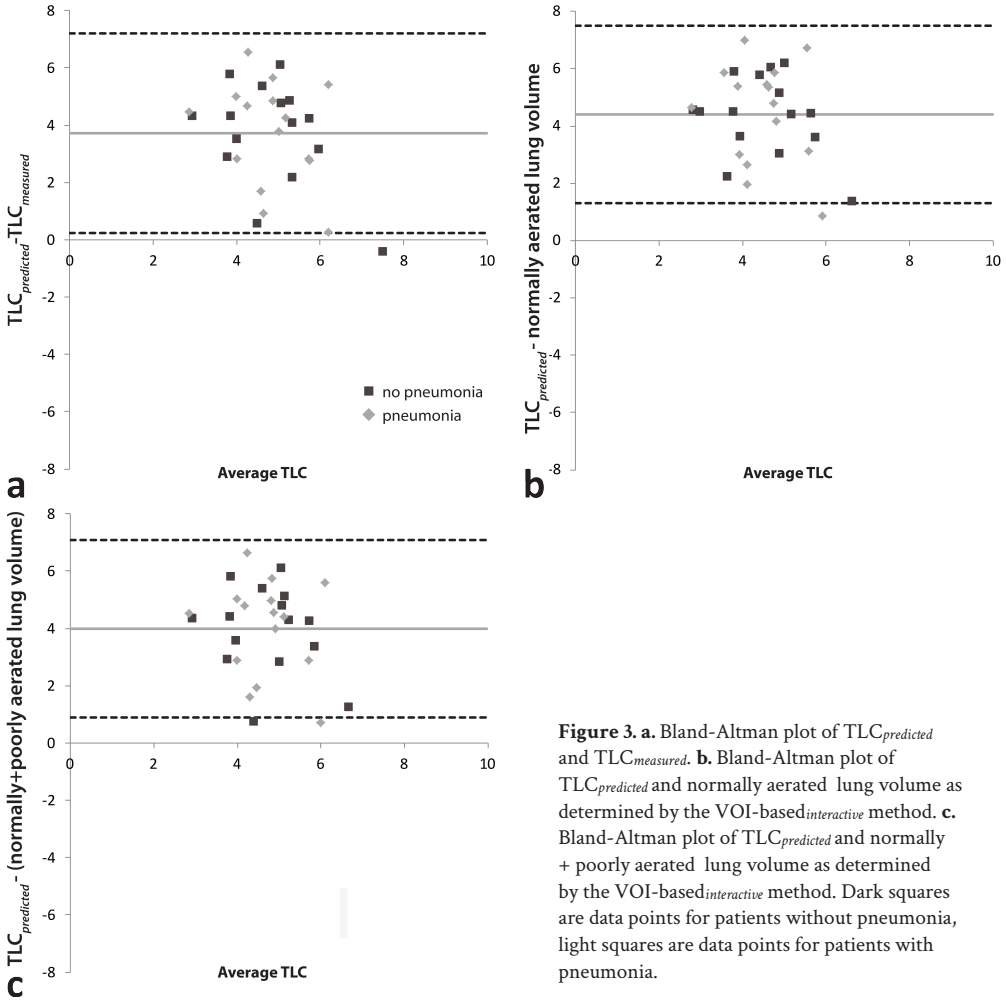


Figure 3. a. Bland-Altman plot of $TLC_{predicted}$ and $TLC_{measured}$. **b.** Bland-Altman plot of $TLC_{predicted}$ and normally aerated lung volume as determined by the VOI-based$_{interactive}$ method. **c.** Bland-Altman plot of $TLC_{predicted}$ and normally + poorly aerated lung volume as determined by the VOI-based$_{interactive}$ method. Dark squares are data points for patients without pneumonia, light squares are data points for patients with pneumonia.

and measured normal lung volume were even larger than the differences between predicted and measured TLC. In Figure 3c, $TLC_{predicted}$ is compared to the sum of normally and poorly aerated lung tissue. Also for this comparison, individual measurements varied substantially from the predictions of TLC based on patient height.

Figure 4 shows an axial slice of one of the CT scans used in this study (a). In panel b, the categorization based on the voxel-based$_{automatic}$ method is shown. The ventral parts of the lungs contained mainly normally aerated tissue, and in the dorsal part of the left lung, a large non-aerated volume was seen. The middle part of the lungs contained poorly aerated tissue. Within these areas, many small islands of abnormally aerated tissue were found. In figure 4c, voxels were labeled based on the VOI-based$_{automatic}$ method. While the global distribution of aeration categories was comparable with the distribution in figure 4b, the

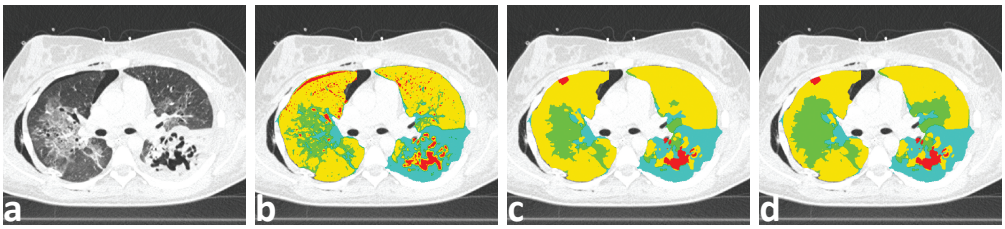


Figure 4. a. Example of an axial slice in one of the CT scans. **b.** Aeration categories based on voxel density value (voxel-based$_{automatic}$). **c.** Aeration categories based on median density value per VOI (VOI-based$_{automatic}$). **d.** Aeration categories based on interactive labeling (VOI-based$_{interactive}$). Red = hyperinflated; yellow: normally aerated; green: poorly aerated; blue: non-aerated.

small islands were absent in this approach. In panel d, results of the VOI-based$_{interactive}$ methods are displayed. Poorly aerated volume increased in comparison with panel c, at the cost of normally aerated volume.

In Figure 5, normally aerated lung volume as determined by all 3 quantification methods are compared. Interactive texture analysis took on average 5 minutes per patient. Panel a compares the results of the voxel-based$_{automatic}$ method to the results of the VOI-based$_{interactive}$ method, and panel b compares the results of the VOI-based$_{automatic}$ method to the results of the VOI-based$_{interactive}$ method. When comparing both panels, the differences between the VOI-based methods were smaller than the differences between the voxel-based$_{automatic}$ and the VOI-based$_{interactive}$ method. For many patients, the VOI-based$_{automatic}$ method yielded a reliable estimate, but in some patients, large adjustments needed to be made, which underlines the added value of an interactive over an automatic method.

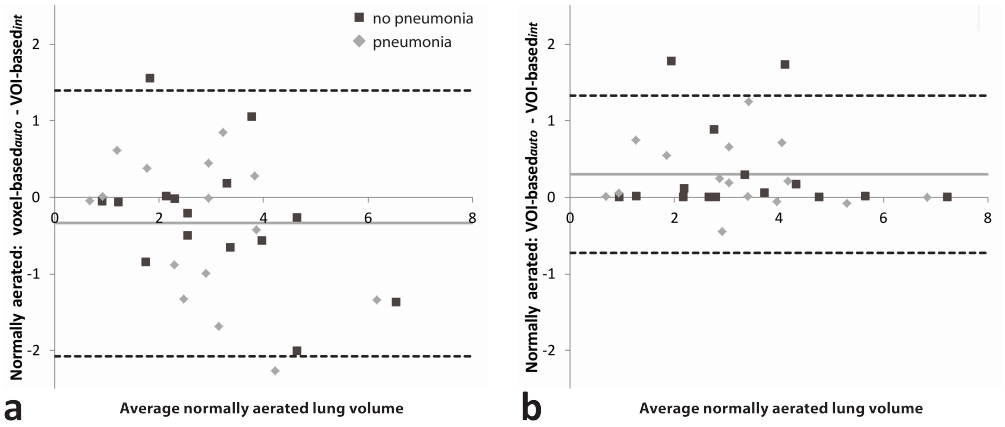


Figure 5. Bland-Altman plots comparing the normally aerated lung volume as determined by **a.** the voxel-based$_{automatic}$ and the VOI-based$_{interactive}$ methods and **b.** the VOI-based$_{automatic}$ and the VOI-based$_{interactive}$ methods. Dark squares are data points for patients without pneumonia, light squares are data points for patients with pneumonia. aut = automatic; int = interactive.

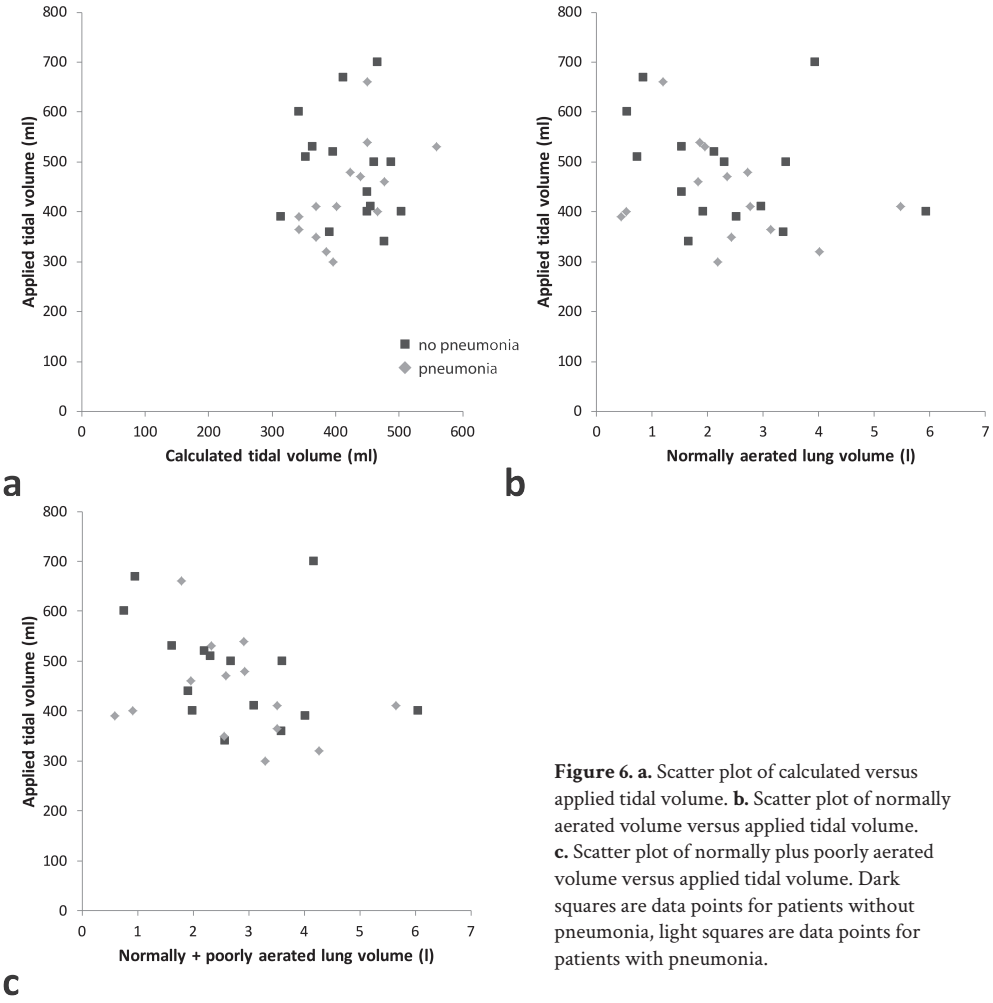


Figure 6. a. Scatter plot of calculated versus applied tidal volume. **b.** Scatter plot of normally aerated volume versus applied tidal volume. **c.** Scatter plot of normally plus poorly aerated volume versus applied tidal volume. Dark squares are data points for patients without pneumonia, light squares are data points for patients with pneumonia.

Figure 6a is a scatter plot of applied versus calculated tidal volume of 6 ml per kg PBW. Overall correlation between the two measures was 0.20, but results varied for patients with and without pneumonia. For patients with pneumonia, the correlation between the two measures was 0.61. Correlation for patients without pneumonia was -0.14.

Figure 6b is a scatter plot of the aerated lung volume, as measured by interactive VOI annotation, versus applied tidal volume. The correlation between aerated lung volume and applied volume was -0.16 (-0.11 for patients with pneumonia and -0.25 for patients without pneumonia). In Figure 6c, the sum of normally and poorly aerated lung volume is plotted against the applied tidal volume. Overall correlation between the two measures was -0.19 (-0.05 for patients with, and -0.35 for patients without pneumonia).

5 Discussion

In this work, we evaluated the use of an interactive segmentation and annotation tool for determining the aeration of lung tissue of patients admitted to the ICU. A reliable estimate of the lung volume that can still perform gas exchange may serve as a first step towards individualized mechanical ventilation settings.

We found that TLC cannot be reliably estimated based on body height in this group. A complication in patients admitted to the ICU is that estimates of lung volume based on body height do not take the volume of damaged lung into account. In clinical practice, this inability to reliably estimate the normally aerated lung volume is supposedly overcome by applying protective mechanical ventilation strategies. However, as far as applied tidal volume is concerned, the factual effect of the protective ventilation depends on the lung volume that is aerated. We show that the normally and poorly aerated lung volumes can be determined using an interactive method, in on average 5 minutes per patient.

In our dataset, no correlation was found between applied V_T and the sum of normally and poorly aerated lung volume. The correlation between applied V_T and normally aerated lung volume was weakly negative. If calculation of the applied tidal volume took the functioning lung volume into account, we would expect this correlation to be positive: the smaller the amount of aerated lung volume, the smaller the V_T that can be applied safely. However, further research is required to determine the optimal relation between normally aerated lung volume and applied V_T.

Using driving pressure as a determinant for ventilator settings may in part overcome this problem, as this will lead to individual settings. Still, since this is an emerging concept, numerous patients are being ventilated with protective ventilation using small tidal volumes. Also, numerous patients are not being ventilated with protective ventilation. It is important to realize that for both two categories aerated lung volumes vary greatly.

While acquiring CT scans of ICU patients may have disadvantages, it can also present benefits, by offering more insight into patient diagnosis and subsequent optimal treatment. Many patients already require a CT scan for diagnostic purposes when admitted to the ICU (Aliaga et al. 2015). These scans could readily be used for determining personalized ventilation settings, and thus require no additional radiation dose.

Existing methods for quantification of normally and abnormally aerated lung tissue have focused on automatic labeling of individual voxels. We have compared the results of a voxel-based method to our VOI-based approach. This VOI-based approach resembles the way

in which human observers would annotate a CT scan: by delineating normal and abnormal areas rather than assigning a label to each individual voxel.

By having an observer interactively refine the automatic VOI-based labeling results, we found that the volume of poorly aerated lung tissue was overestimated by the VOI-based$_{automatic}$ method in some patients. More advanced methods to distinguish different aeration categories might reduce this overestimation, since the threshold-based method is not robust against factors such as respiration depth and image acquisition parameters, both of which can influence the measured densities. In an interactive quantification system, the effects of these factors can be corrected by the observer. An alternative interactive approach would be to let observers vary the thresholds between the different aeration categories using a slider, which could increase annotation speed as well, as observers would not have to adjust the label of individual VOIs.

A limitation of the work presented in this paper can be seen in figure 4. The borders between consolidated lung and surrounding tissue are not always sharply defined, since their density values are similar. When VOIs are made, they may run across lung borders, resulting in irregular segmentation edges. This can for example be remedied by allowing observers to adapt the predefined VOIs manually, or by adding a post-processing step to the segmentation algorithm, such as smoothing. On the other hand, it should be noted that the differences between $TLC_{measured}$ and $TLC_{predicted}$ are large. Therefore, deviations of volume measurements in the milliliter range are not problematic and precision at voxel level is less relevant for this application.

Although we have shown that calculation of the normally and poorly aerated lung volume is possible, the relation between normally and poorly aerated lung volume and ideal tidal volume remains to be elucidated. Future research should be directed at finding the relation between normally and poorly aerated lung volume and optimal tidal volume settings.

6 Conclusion

In our study group of ICU patients, we were able to determine the volume of hyperinflated, normally aerated, poorly aerated, and non-aerated lung tissue. In contrast to automatic quantification methods, a human observer reviews and corrects automatic quantification results, to ensure that they reflect the degree of injury to the lungs. Although the described study is retrospective and we cannot draw any conclusions with respect to whether personalized settings would perform better in clinical practice, the presented study is a first step towards a system in which aerated lung tissue can be quantified and taken into account when determining mechanical ventilation settings.

References

Aliaga M, Forel JM, De Bourmont S, Jung B, Thomas G, Mahul M, Bisbal M, Nougaret S, Hraiech S, Roch A, Chaumoitre K, Jaber S, Gainnier M, Papazian L. Diagnostic yield and safety of CT scans in ICU. *Intensive Care Med.* **2015**;41(3):436-43.

Amato MB, Meade MO, Slutsky AS, Brochard L, Costa EL, Schoenfeld DA, Stewart TE, Briel M, Talmor D, Mercat A, Richard JC, Carvalho CR, Brower RG. Driving pressure and survival in the acute respiratory distress syndrome. *N Engl J Med.* **2015**;372(8):747-55.

American Thoracic Society, European Society of Intensive Care Medicine, Societé de Réanimation Langue Française. International consensus conferences in intensive care medicine. Ventilator-associated lung injury in ARDS. *Intensive Care Med.* **1999**;25(12):1444-52.

Devine BJ. Gentamicin therapy. *Drug Intell Clin Pharm.* **1974**;8:650-5

Caironi P, Cressoni M, Chiumello D, Ranieri M, Quintel M, Russo SG, Cornejo R, Bugedo G, Carlesso E, Russo R, Caspani L, Gattinoni L. Lung opening and closing during ventilation of acute respiratory distress syndrome. *Am J Respir Crit Care Med.* **2010**;181(6):578-86.

Fanelli V, Ranieri VM. Mechanisms and clinical consequences of acute lung injury. *Ann Am Thorac Soc.* **2015**;12 Suppl 1:S3-8.

Gattinoni L, Caironi P, Pelosi P, Goodman LR. What has computed tomography taught us about the acute respiratory distress syndrome? *Am J Respir Crit Care Med.* **2001**;164(9):1701-11.

Gattinoni L1, Chiumello D, Cressoni M, Valenza F. Pulmonary computed tomography and adult respiratory distress syndrome. *Swiss Med Wkly.* **2005**;135(11-12):169-74.

Gattinoni L, Caironi P, Valenza F, Carlesso E. The role of CT-scan studies for the diagnosis and therapy of acute respiratory distress syndrome. *Clin Chest Med.* **2006**;27(4):559-70; abstract vii.

Gattinoni L, Cressoni M. Quantitative CT in ARDS: towards a clinical tool? *Intensive Care Med.* **2010**;36(11):1803-4.

Kockelkorn TTJP, Schaefer-Prokop CM, Bozovic G, Muñoz-Barrutia A, van Rikxoort EM, Brown MS, de Jong PA, Viergever MA, van Ginneken B. *Interactive lung segmentation in abnormal human and animal chest CT scans.* Med Phys. **2014**;41(8):081915.

Kockelkorn TTJP, de Jong PA, Schaefer-Prokop CM, Wittenberg R, Tiehuis AM, Gietema HA, Grutters JC, Viergever MA, van Ginneken B. *Semi-automatic classification of textures in thoracic CT scans. Phys Med Biol.* **2016**;61(16):5906-5924.

Malbouisson LM, Muller JC, Constantin JM, Lu Q, Puybasset L, Rouby JJ; CT Scan ARDS Study Group. Computed tomography assessment of positive end-expiratory pressure-induced alveolar recruitment in patients with acute respiratory distress syndrome. *Am J Respir Crit Care Med.* **2001**;163(6):1444-50.

Patroniti N, Bellani G, Maggioni E, Manfio A, Marcora B, Pesenti A. Measurement of pulmonary edema in patients with acute respiratory distress syndrome. *Crit Care Med.* **2005**;33(11):2547-54.

Puybasset L, Cluzel P, Gusman P, Grenier P, Preteux F, Rouby JJ. Regional distribution of gas and tissue in acute respiratory distress syndrome. I. Consequences for lung morphology. CT Scan ARDS Study Group. *Intensive Care Med.* **2000**;26(7):857-69.

Quanjer PH, Tammeling GJ, Cotes JE, Pedersen OF, Peslin R, Yernault JC. Lung volumes and forced ventilatory flows. *Eur Respir J.* **1993**;6 Suppl 16:5-40.

Reske AW, Reske AP, Gast HA, Seiwerts M, Beda A, Gottschaldt U, Josten C, Schreiter D, Heller N, Wrigge H, Amato MB. Extrapolation from ten sections can make CT-based quantification of lung aeration more practicable. *Intensive Care Med.* **2010**;36(11):1836-44.

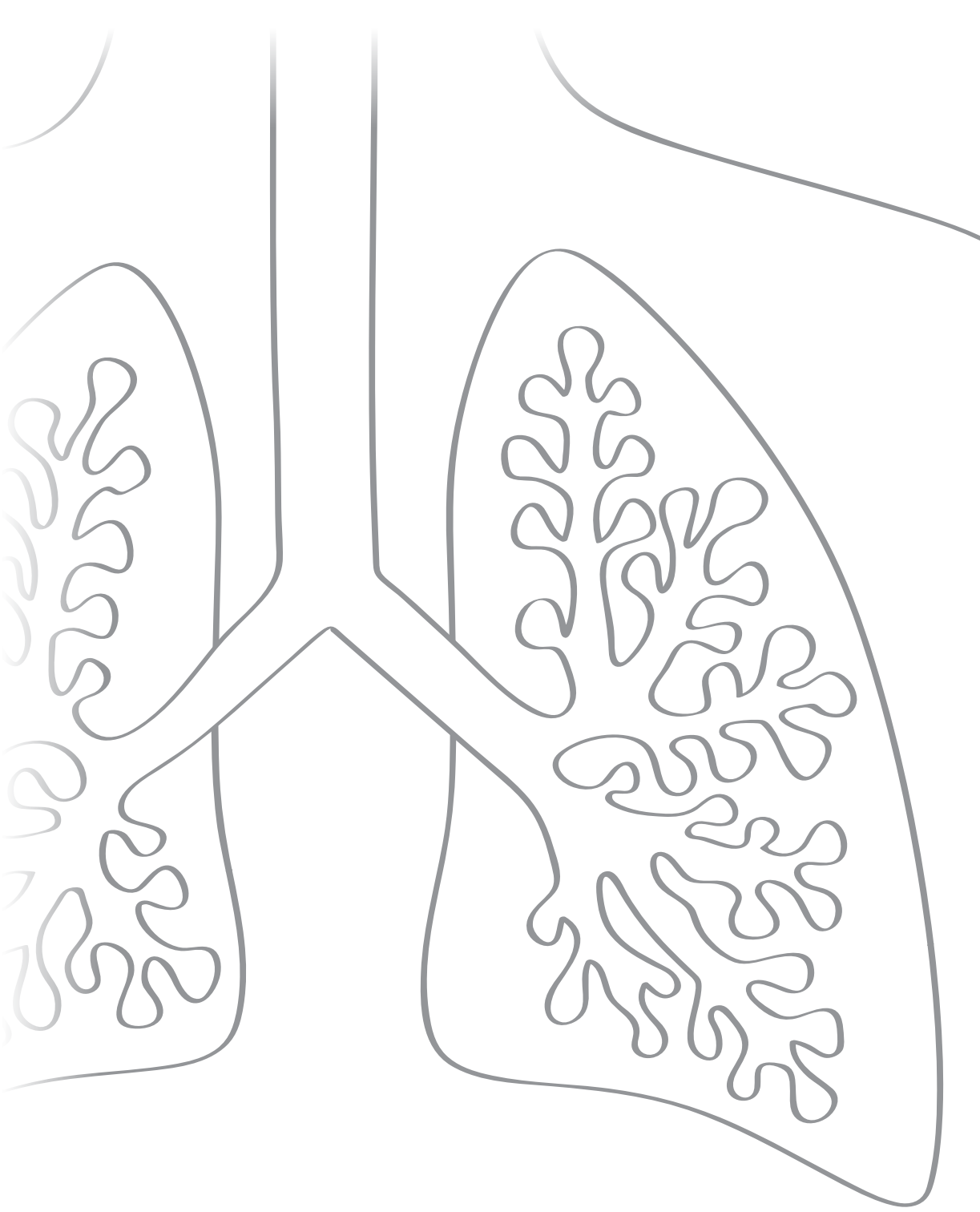
Rouby JJ, Puybasset L, Nieszkowska A, Lu Q. Acute respiratory distress syndrome: lessons from computed tomography of the whole lung. *Crit Care Med.* **2003**;31(4 Suppl):S285-95.

Rylander C, Högman M, Perchiazzi G, Magnusson A, Hedenstierna G. Oleic acid lung injury: a morphometric analysis using computed tomography. *Acta Anaesthesiol Scand.* **2004**;48(9):1123-9.

Schreiter D, Reske A, Stichert B, Seiwerts M, Bohm SH, Kloeppel R, Josten C. Alveolar recruitment in combination with sufficient positive end-expiratory pressure increases oxygenation and lung aeration in patients with severe chest trauma. *Crit Care Med.* **2004**;32(4):968-75.

Slutsky AS, Ranieri VM. Ventilator-induced lung injury. *N Engl J Med.* **2014**;370(10):980

The Acute Respiratory Distress Syndrome Network. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N Engl J Med.* **2000**;342(18):1301-8.



Summary & general discussion

Summary

This thesis describes and evaluates an interactive annotation system for chest CT scans. As a first step, the structure of interest is segmented automatically. This structure is then divided into smaller volumes of interest (VOIs) containing one type of texture. These VOIs are automatically labeled, either by a classifier or using a heuristic approach.

An observer interactively refines the automatically assigned labels. This can be done by changing the label of incorrectly labeled VOIs one by one, or by training a classifier. In the latter case, the observer is shown an axial slice of the scan, with a semi-transparent overlay of the automatic labeling results of all VOIs intersecting with that slice. The observer corrects the labeling errors and the VOIs reviewed by the observer are used to train a classifier. This classifier classifies the VOIs in a second axial slice and the observer again corrects any mistakes. The VOIs are added to the training data and the classifier is retrained. In this way, the classifier is iteratively trained to perform the classification task at hand for the scan at hand, adapted to the preferences of the observer. The process of correction, retraining, and classification is ended when a predefined percentage of the structure of interest is annotated, or on user prompt. After that, the classifier is trained again and all remaining VOIs are classified. The observer inspects all annotations to correct any remaining errors, after which he/she saves the annotation results.

In **Chapter 2,** interactive classification was applied to CT scans of the lungs of interstitial lung disease (ILD) patients. The aim was to develop a tool for annotating normal and seven categories of abnormal lung tissue - decreased density, consolidation, honeycombing, ground glass, crazy paving, NSIP pattern, and nodular pattern. In this chapter, a simulation approach was used to compare different interactive annotation strategies. First, we considered an approach in which all VOIs were labeled automatically by a classifier trained on VOIs from other, previously annotated scans. In this scenario, the labels of on average 58% of the VOIs needed to be changed by the observer. In a second approach, annotations were made interactively. The VOIs in the first slice were labeled using a heuristic approach: they were labeled as normal tissue, since that was the texture class with the highest prior probability in our data. After that, slices were annotated interactively. A classifier was trained on the VOIs in the slices already reviewed by the observer. When 50% of the VOIs were labeled interactively, the remainder of the scan was labeled automatically by the interactively trained classifier. Using this approach, the label of on average 21% of all VOIs needed to be corrected. The last approach was similar to the second, except for labeling of the first slice, which was done by a classifier trained on VOIs from previously annotated scans. On average, the labels of 20% of all VOIs needed to be changed.

In **Chapter 3**, optimization strategies for the interactive annotation framework were investigated. Lungs were segmented and divided into VOIs. The aim was to decrease the number of VOI labels that observer would have to label or correct manually in order to annotate all lung voxels in chest CT scans of ILD patients. Eight types of textures were distinguished: normal tissue, decreased density, consolidation, honeycombing, ground glass, crazy paving, NSIP pattern, and nodular pattern. Simulation software was used to test several scenarios. First, automatic classification experiments were performed to compare different manners in which training data derived from previously annotated scans could be used during interactive annotation. The performance of four classifiers was compared: a classifier trained on annotations from one observer, a classifier trained on annotations from three observers, a classifier trained on consensus annotations, and an ensemble of classifiers. Each classifier in the ensemble was trained on data from a different source. All experiments were conducted twice, once without and once with texture selection. Without texture selection, training data from all 8 textures was included. With texture selection, only training data from the texture classes that were present in the scan under consideration was used. In a second set of experiments, interactive annotation was simulated to evaluate the effects of (1) texture selection, (2) the use of a classifier trained on previous annotations, and (3) offering observers a choice between different interactive or automatic classification results as a starting point for correction. For all experiments in these two sets, classification accuracies were calculated. The best performing protocol was the one in which observers performed texture selection and in which they could choose which classification results they wanted to correct. A median accuracy of 88% was reached. The final set of experiments evaluated four strategies for selection of the axial slices that were presented to the observer. Choosing slices of which the classifier was less certain resulted in a lower number of slices that observers had to inspect than choosing slices at random. Dividing the scan into 5 equally large parts in axial direction, from which slices were chosen alternatively, did not yield an advantage as compared with a scenario in which slices could be chosen from the entire scan, neither in terms of the number of slices that had to be inspected, nor in terms of classification accuracy.

Chapter 4 discussed the application of interactive annotation to lung segmentation. Thirty-two CT scans were selected: 8 from patients who were to undergo lung transplantation, 8 from the LOLA11 challenge, 8 from a pig model and 8 murine micro-CT scans. Chests were segmented automatically and divided into VOIs. VOIs with an average CT number of -500 or below were classified as lung tissue, the remainder as non-lung tissue. A human observer interactively refined these segmentations, either by a supervised approach in which a classifier was trained interactively in a slice-by-slice manner, or by correcting

all automatically generated labels. Results were compared to manual delineations of the lungs. Supervised interactive lung segmentation took under 9 min of user interaction on average. The labels of on average 2.0% of all VOIs in a scan had to be changed. Lung segmentation without the supervised approach took 13 min on average, for relabeling on average 3.0% of all VOIs. Interactively obtained segmentations corresponded well to manual delineations, as evaluated in eight axial slices per scan. The average Dice similarity coefficient was 0.933.

Finally, **Chapter 5** discussed the application of interactive lung segmentation, followed by texture analysis in 30 CT scans of intensive care unit (ICU) patients who required mechanical ventilation. While mechanical ventilation is a potentially life-saving intervention, it may cause additional lung damage in patients with acute respiratory distress syndrome (ARDS) when too large tidal volumes (V_T) are applied. In clinical practice, V_T is determined based on body height and the presence or absence of ARDS. The aim of this chapter was to determine total lung capacity and the volume of normally aerated lung tissue based on CT scans, as a first step towards personalized calculation of the optimal V_T. Lungs were segmented interactively using the framework described in Chapter 4. Then, lungs were divided into smaller VOIs and labeled as hyperinflated, normally aerated, poorly aerated, and non-aerated based on their median CT value. These labels were interactively refined. We found that TLC and normally aerated lung volume could not be reliably estimated using only body height and the presence or absence of ARDS. Using the described interactive approach, reliable estimates can be made for each patient individually.

General discussion

In the following paragraphs, limitations and future perspectives of the different applications of interactive texture analysis described in this thesis will be discussed.

Limitations

Although the results of interactive annotation and segmentation as described in Chapters 2-5 are promising, there are a number of amendments that can be made.

First, the method of VOI creation can be optimized. In the current implementation, VOIs do not always precisely follow the boundaries between two textures. In the creation process, seeds are selected, to which all voxels within a user-defined distance are added to form the initial VOI. If seeds are selected close to boundaries between two textures, the growing VOI will contain more than one type of texture, resulting in incorrect segmentations. An alternative method for VOI creation may solve this problem, for example by dividing the structure of interest into cubes, of which the boundaries are iteratively refined. In this iterative refinement, texture homogeneity measures can be used to determine if voxels neighboring VOI boundaries should remain in their original VOI, or if they belong on the other side of the boundary. Finally, superpixels or supervoxels have been used by other groups for segmentation and anomaly detection purposes in various imaging modalities and organs (e.g. Chu et al 2015, Bejnordi et al 2016, Liao et al 2016). Superpixels are groups of pixels with similar color and brightness values; supervoxels are their 3d counterparts. Supervoxels could serve as an alternative for the VOIs described in this thesis.

Second, in Chapters 2 and 3 we used simulation software to mimic observer behavior. In order to assess the usability of the proposed interactive annotation system, experiments should be repeated with human observers. These observers can also provide feedback on other aspects of interactive annotation besides classification accuracy. In addition, the behavior of the simulated observers was not influenced by interactive classification results. As interobserver variability is substantial in ILD annotation, it would be interesting to evaluate how observers are influenced by the classification results that are shown. If observers are asked to annotate scans twice, once by manually labeling all VOIs, and once interactively, interactive annotations can be compared to manual annotations on the one hand, and to interactive classification results on the other hand. From these comparisons, the degree to which observers are influenced by algorithm output can be assessed. It should be noted that observer behavior being influenced by classification results is not necessarily negative. If this influence leads to more objective and consistent annotation behavior, this would be an additional advantage of interactive over manual annotation.

Finally, only a limited number of features and classifiers were evaluated for optimization of the classification protocol. An important requirement for a classifier used in interactive classification is that it can be trained fast, using a training dataset of limited size. Other groups have reported on the use of deep learning techniques for classification of ILD textures (Gao et al. 2015; Anthimopoulos et al. 2016). Deep learning is inspired by human learning strategies. Instead of being given features that are selected by humans, deep learning algorithms themselves determine which features best represent the data. These features are organized in a hierarchical manner, the higher level features being derived from the more abstract lower level features. In general, deep learning requires a considerable amount of training data and training time. Within the interactive classification framework as proposed in chapters 2 and 3, it could be used for classification of the first axial slice, or possibly slices, when the interactive classifier is relatively untrained. Part of this pre-trained network can also be used for feature extraction for the remainder of the classification task. As there is considerable overlap between the dataset used for the pre-trained network and the dataset obtained from the scan under consideration, fine-tuning the algorithm to the scan under consideration can be done efficiently by retraining only the top layers (Yosinski et al 2014).

Future perspectives

Interactive segmentation

Segmentation is one of the first components in many CAD algorithms. In normal lungs, successful automatic segmentation protocols have been developed (van Rikxoort and van Ginneken 2013). However, for many abnormal cases automatic segmentation is more complicated. Interactive segmentation approaches for segmentation of medical images have been described (Zhao and Xie 2013), and recently, Lassen-Schmidt and coworkers presented an interactive approach for segmentation of pulmonary lobes (2015).

The main strength of the our interactive segmentation framework is that it is capable of segmentation of abnormal human lungs, but that it can also be used without further amendments for segmentation of porcine lung in CT scans, and even murine lungs in micro-CT scans. With further modifications, VOI-based segmentation could also be applied to segmentation of other structures and to other imaging modalities. Depending on the nature of the intended application, changes could for example be made to the process of VOI formation to better capture the boundaries of the structure that should be segmented, as discussed in Chapter 4. Another amendment could be to couple this interactive segmentation framework to an automatic one. One of the disadvantages of VOI based seg-

mentation as described in this thesis is that the method sometimes produces VOIs that do not follow the boundaries of the segmented structure precisely, in areas where automatic methods are able to produce correct results. By taking the result of an automatic segmentation method into account when calculating the VOIs, both methods can complement each other: automatic segmentation produces smooth results in the easier areas, whereas interactive segmentation can be used only for the more difficult parts.

Interactive texture analysis

Automatic texture analysis of chest CT scans of patients with suspected ILD can in theory assist physicians in two ways. First, analysis of the types of lung textures, their quantities, and their distribution throughout the lungs can help in making the correct diagnosis. Second, by measuring qualitative and quantitative changes in textures over time, disease progression and treatment response can be monitored. For automatic texture analysis to become feasible, ground truth annotations are necessary for training such a system. The interactive annotation framework described in Chapters 2 and 3 was developed for this purpose. Whereas the topic of automatic ILD texture classification in chest CT scans has attracted considerable attention, the process of obtaining reliable annotations is not often discussed. The substantial amount of inter-observer variability in Chapters 2 and 3 of this thesis underlines the importance of such a discussion.

The first and most important step is definition of the texture patterns that should be annotated. In literature, many different, but partially overlapping categorizations have been proposed (Sluimer et al 2006, Depeursinge et al. 2012, Kockelkorn et al. 2016). Using one list of texture patterns for all ILD-related research would allow for an easier exchange of datasets. As a starting point, four main patterns can be used: nodular patterns, reticular patterns, increased density, and decreased density (see for example Schaefer-Prokop 2014), all of which can be further subdivided. The list of textures should be mutually exclusive and collectively exhaustive, in order to cover all possible textures. If consensus on the texture categories is reached, annotations can be made interactively in a research setting. Since interactive annotation is fast, each scan can be annotated by a number of experts. By comparing the texture annotations of different observers, a database containing consensus readings can be made.

In Chapter 3, experiments were conducted to see if training data from previously annotated scans can be used to classify VOIs when little or no interactive training data is available. For the limited number of scans included in this study, this was not the case. It would be interesting to repeat those experiments using a larger training dataset with consensus

readings. In addition, the use of more sophisticated classification techniques, such as deep learning all may help to increase classification accuracy.

Once a database of textures has been established and a classification system based on these textures is developed, they can be coupled to the interactive annotation environment. In this way, clinicians can use it as a reference standard when assessing lung texture patterns in patients with ILD. In addition, tests can be done to evaluate the use of interactive annotation for quantification of the different ILD texture patterns (Kim et al 2015, Jacob et al 2016, Nakagawa et al 2016). The scans annotated in this way can be added to the texture database, along with information on the patients and their diagnosis. Meanwhile, the texture database can be used in research as well, for example for comparison of different classification algorithms in a grand challenge.

When enough scans of patients with different diagnoses have been gathered, research can focus on predicting diagnosis based on textures present in unseen scans. As an accurate diagnosis of an individual ILD requires a multidisciplinary approach, other clinical datasources such as anamnesis, results of blood tests, and pulmonary function testing can be incorporated (Meyer 2014, Wallis and Spinks 2015). Initial experiments can focus on a limited number of diseases with the highest incidence; later on more rare diseases can be added. The disease database can be integrated into the interactive annotation environment. Now interactive annotation can be used to help radiologists not only in quantification, but also in making the correct diagnosis. In addition, it can be used for education purposes. Individual ILDs being rare, a reference database can help starting radiologists become acquainted with the different radiological appearances and individual diseases.

Calculation of aerated lung volume in ICU patients

The topic covered in Chapter 5, measurement of aerated lung volume in ICU patients, is an example of a practical clinical application for which interactive texture annotation can be used. As a next step, a larger retrospective study can be performed. In this study, aerated lung volume is calculated based on CT scans as described in Chapter 5 of this thesis and can then be related to the applied tidal volume (V_T). Patients can be divided into groups reflecting this relation, for example into large applied V_T with respect to measured aerated volume, normal V_T with respect to measured aerated volume, and small V_T with respect to measured aerated volume. In the next step, a risk assessment of adverse events, such as ventilator-associated lung injury (VALI), can be made for each of these groups. If patients in one of the groups are significantly more at risk for developing an adverse event, this is an indication that the proposed method for determining aerated volume can be useful for calculation of personalized V_T.

References

Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Trans Med Imaging* **2016** doi:10.1109/TMI.2016.2535865.

Chu J, Min H, Liu L, Lu W. A novel computer aided breast mass detection scheme based on morphological enhancement and SLIC superpixel segmentation. *Med Phys.* **2015**;42(7):3859-69.

Bejnordi BE, Balkenhol M, Litjens G, Holland R, Bult P, Karssemeijer N, van der Laak JA. Automated detection of DCIS in whole-slide H&E stained breast histopathology images. *IEEE Trans Med Imaging*. **2016**. DOI: 10.1109/TMI.2016.2550620.

Depeursinge A, Van de Ville D, Platon A, Geissbuhler A, Poletti P-A, Müller H. Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames. *IEEE Transactions on Information Technology in Biomedicine* **2012;**16(4): 665-675.

Gao M, Bagci U, Lu L, Wu A, Buty M, Shin H, Roth H, Papadakis GZ, Depeursinge A, Summers RM, Xu Z, Mollura DJ. Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* **2016** doi: 10.1080/21681163.2015.1124249

Jacob J, Bartholmai BJ, Rajagopalan S, Kokosi M, Nair A, Karwoski R, Raghunath SM, Walsh SL, Wells AU, Hansell DM. Automated Quantitative Computed Tomography Versus Visual Computed Tomography Scoring in Idiopathic Pulmonary Fibrosis: Validation Against Pulmonary Function. *J Thorac Imaging*. **2016**;31(5):304-11.

Kim HJ, Brown MS, Chong D, Gjertson DW, Lu P, Kim HJ, Coy H, Goldin JG. Comparison of the quantitative CT imaging biomarkers of idiopathic pulmonary fibrosis at baseline and early change with an interval of 7 months. *Acad Radiol.* **2015**;22(1):70-80.

Kockelkorn TTJP, de Jong PA, Schaefer-Prokop CM, Wittenberg R, Tiehuis AM, Gietema HA, Grutters JC, Viergever MA, van Ginneken B. Semi-automatic classification of textures in thoracic CT scans. *Phys Med Biol.* **2016**;61(16):5906-24.

Lassen-Schmidt BC. Interactive Segmentation of the Pulmonary Lobes. In: *Automatic and Interactive Segmentation of Pulmonary Lobes and Nodules in Chest CT Images*. **2015** (PhD thesis). Radboud Universiteit Nijmegen.

Liao X, Zhao J, Jiao C, Lei L, Qiang Y, Cui Q. A Segmentation Method for Lung Parenchyma Image Sequences Based on Superpixels and a Self-Generating Neural Forest. *PLoS One.* **2016**;11(8):e0160556.

Meyer KC. Diagnosis and management of interstitial lung disease. *Transl Respir Med.* **2014**;2:4.

Nakagawa H, Nagatani Y, Takahashi M, Ogawa E, Tho NV, Ryujin Y, Nagao T, Nakano Y. Quantitative CT analysis of honeycombing area in idiopathic pulmonary fibrosis: Correlations with pulmonary function tests. *Eur J Radiol.* **2016**;85(1):125-30.

van Rikxoort EM, van Ginneken B. Automated segmentation of pulmonary structures in thoracic computed tomography scans: a review. *Phys Med Biol.* **2013**;58(17):R187-220.

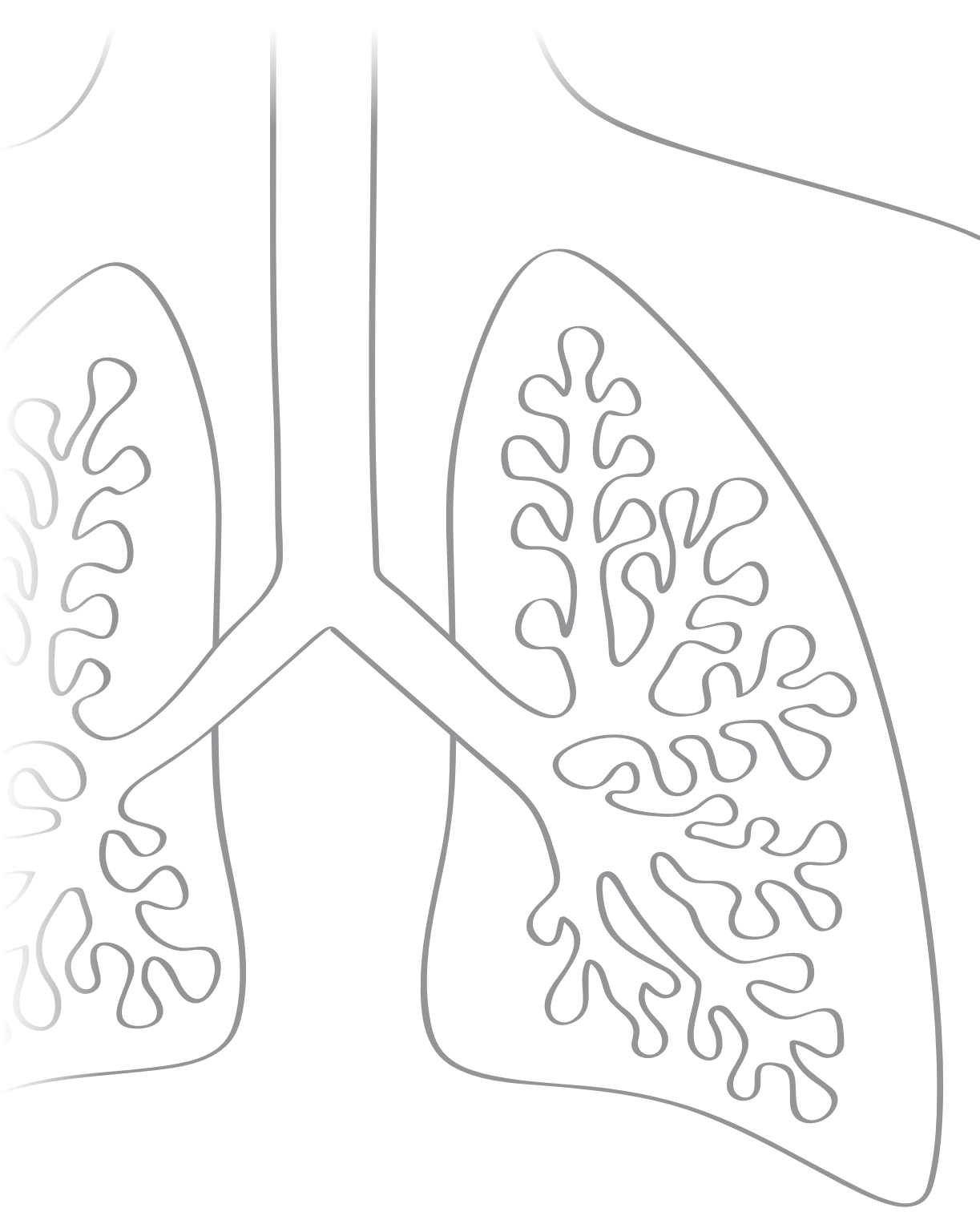
Schaefer-Prokop C. HRCT patterns of the most important interstitial lung diseases. *Radiologe.* **2014**;54(12):1170-9. doi: 10.1007/s00117-014-2734-3.

Sluimer IC, Prokop M, Hartmann I, van Ginneken B. Automated classification of hyperlucency, fibrosis, ground glass, solid, and focal lesions in high-resolution CT of the lung. *Med Phys.* **2006**;33(7):2610-20.

Wallis A, Spinks K. The diagnosis and management of interstitial lung diseases. *BMJ.* **2015**;350:h2072.

Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems 27* **2014:**3320-3328

Zhao F, Xie X. Overview on interactive medical segmentation. *Annals of the BMVA.* **2013**:2013(7);1–22.



Addenda

Samenvatting in het Nederlands

Introductie

Medische beeldvorming stelt artsen in staat de binnenzijde van het menselijk lichaam te onderzoeken, op een manier die niet of weinig belastend is voor de patiënt. Verschillende vraagstukken vragen om verschillende beeldvormende technieken. Voor gedetailleerd onderzoek van de longen is computertomografie (CT) de aangewezen manier. Dit proefschrift behandelt interactieve computergestuurde methoden, die zijn ontwikkeld om radiologen te helpen bij de analyse van CT-scans van de thorax. De beschreven methoden zijn gebruikt voor interactieve segmentatie van de longen en voor het classificeren van normaal en abnormaal longweefsel in CT-scans van patiënten met interstitiële longziekten.

Deze samenvatting introduceert een aantal belangrijke begrippen uit de bovenstaande paragraaf en vat hoofdstuk 2 tot en met 5 samen.

Computertomografie

Sinds haar introductie in de jaren 70 van de vorige eeuw is CT één van de meest gebruikte medische beeldvormingstechnieken. Bij een CT-scan maakt een röntgenapparaat een serie foto's, steeds uit een iets andere positie. Een computer voegt al deze opnames samen, zodat er een driedimensionaal beeld ontstaat. Dat beeld is opgebouwd uit zogenaamde voxels. Elke voxel heeft een waarde, uitgedrukt in Hounsfield eenheden (HU), die een indicatie geeft van de dichtheid van de materie op die plek. Longen kunnen goed worden gevisualiseerd met CT-scans. De lucht in de longen heeft een lage dichtheid, waardoor de longen zichtbaar zijn als donkere structuren, waarin de bloedvaten en de wanden van de grotere luchtwegen te zien zijn als lichte structuren. In 2013 werden er in Nederland 1,3 miljoen CT-scans gemaakt, een verdrievoudiging sinds het begin van de jaren 90. Tussen de 20% en 25% van deze scans waren CT-scans van de thorax. Deze toename is te verklaren door de vooruitgang in de geneeskunde in het algemeen en in de medische beeldvorming in het bijzonder. Daarnaast kunnen vergrijzing en mogelijk ook toegenomen patiëntenemancipatie een rol spelen.

Hoewel de voordelen van de toegenomen beschikbaarheid van medische beeldvormende technieken duidelijk zijn, zit er ook een keerzijde aan. Een voor de hand liggend neveneffect van CT-scans is de verhoogde blootstelling aan Röntgenstraling, wat kan leiden tot schade aan het DNA en eventueel tot kanker. Daarnaast zorgt het maken van meer beelden tot een toename van zorgkosten. Een minder vanzelfsprekend gevolg is een toenemende

behoefte aan radiologen voor het interpreteren van de beelden. Om dat laatste probleem op te lossen kunnen uiteraard meer radiologen worden opgeleid, maar een alternatieve aanpak is het ontwikkelen van automatische interpretatiemethoden, zodat computers een deel van het werk van radiologen over kunnen nemen.

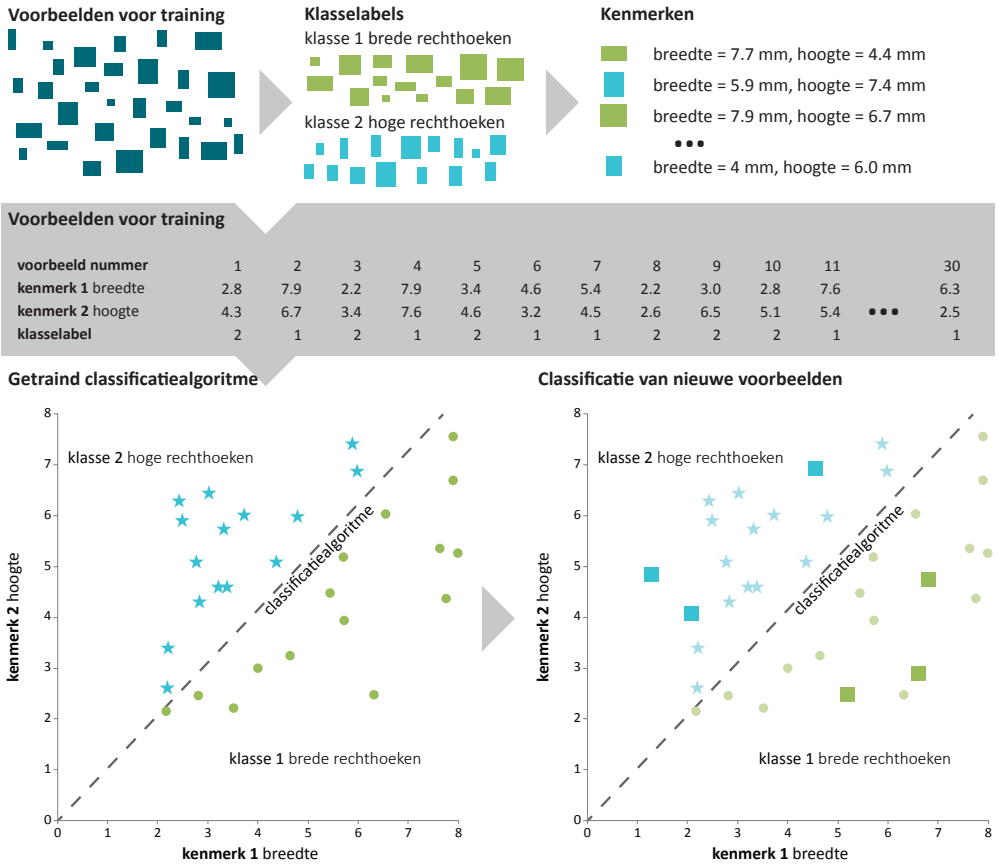
Computergestuurde diagnostiek

Voor veel vakgebieden bestaat automatisering uit het expliciet programmeren van machines om bepaalde taken uit te voeren. Een voorbeeld hiervan is een vaatwasser, die zijn werk kan doen op basis van een set instructies: wat hij moet doen is duidelijk gedefinieerd en altijd hetzelfde. Voor medische beeldanalyse ligt dat wat ingewikkelder. Een computerprogramma dat is ontworpen om afwijkingen op te sporen kun je niet expliciet vertellen waar hij die afwijkingen kan vinden, omdat elke patiënt, elke scan en elke abnormaliteit anders is. Zo'n programma moet in staat zijn te leren van voorbeelden, op een manier die vergelijkbaar is met de manier waarop mensen leren. Onderzoekers op het gebied van computergestuurde diagnostiek (computer-aided diagnosis of CAD) houden zich hiermee bezig. Het meest succesvolle voorbeeld van CAD is de automatische detectie van borstkanker in mammogrammen, die in de Verenigde Staten al op grote schaal wordt gebruikt.

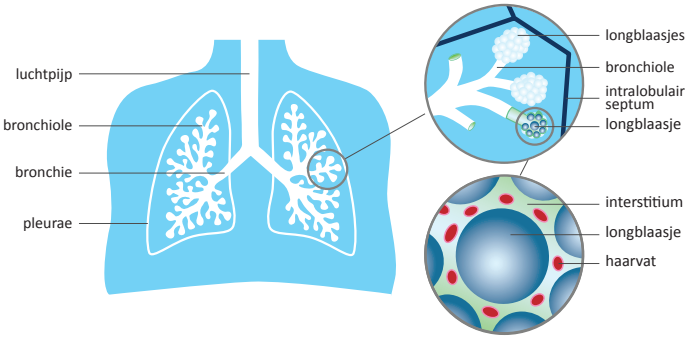
Classificatie

Automatische classificatie is een belangrijk onderdeel van computergestuurde diagnostiek: door middel van voorbeeldobjecten leert een classificatiealgoritme bepaalde structuren of afwijkingen te herkennen. Bij dit leerproces is de keuze van geschikte voorbeelden van groot belang. Deze voorbeelden bevatten twee elementen: een numerieke beschrijving van hun eigenschappen, de kenmerken, en een klasselabel, dat aangeeft tot welke categorie het voorbeeld behoort. Aan de hand van de gelabelde voorbeelden leert het algoritme de relatie tussen de kenmerken en labels en kan het de labels van nieuwe objecten op basis van hun kenmerken voorspellen (zie figuur 1).

De voorbeelden waarmee een classificatiealgoritme getraind is, hebben invloed op de mate waarin het algoritme in staat is de labels van nieuwe objecten correct te voorspellen. Een expert bepaalt de labels van de voorbeelden. Hiermee definieert hij of zij de gouden standaard voor het classificatiealgoritme: als de voorbeelden verkeerd gelabeld zijn, kan het classificatiealgoritme onjuiste verbanden tussen kenmerken en labels leren. Het verkrijgen van een betrouwbare verzameling van gelabelde voorbeelden is daarom cruciaal, maar zeker niet triviaal bij het maken van een CAD-algoritme.



Figuur 1. Schematisch overzicht van automatische classificatie, waarbij een classificatiealgoritme hoge en brede rechthoeken leert onderscheiden. Een expert labelt een aantal voorbeelden als 'breed' of 'hoog'. Voor ieder voorbeeld zijn de breedte en de hoogte van de rechthoek gemeten. De relatie tussen de kenmerken en de labels is weergegeven in het spreidingsdiagram linksonder. Hier is het classificatiealgoritme een rechte lijn. Alle rechthoeken boven de lijn $y=x$ zijn hoog, alle rechthoeken onder de lijn zijn breed. Nu het algoritme getraind is, kan het ook nieuwe rechthoeken classsificeren als 'breed' of 'hoog', op basis van hun kenmerken.



Figuur 2. Schematisch overzicht van de longen, bronchiolen en longblaasjes.

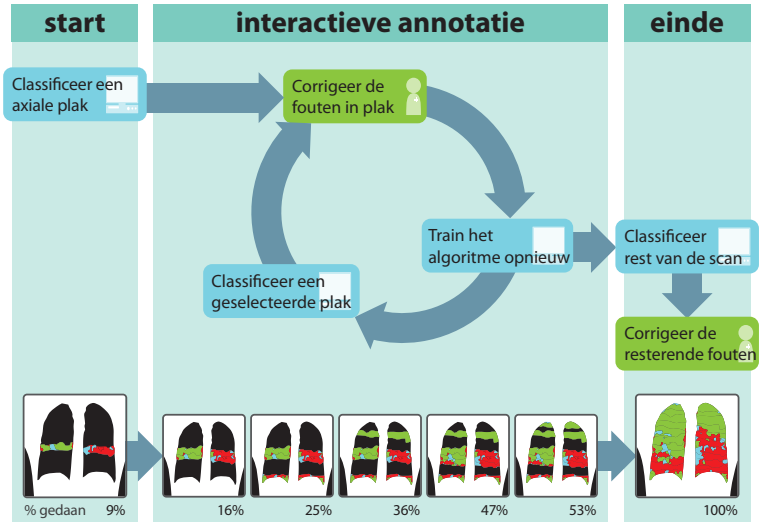
In het onderzoek beschreven in dit proefschrift zijn verschillende classificatiealgoritmes gebruikt. Hoewel de principes hetzelfde zijn als bij het voorbeeld in figuur 1, is de complexiteit van de beschreven problemen veel groter. Daarom maken de algoritmes gebruik van veel meer dan twee kenmerken en zijn ze minder eenvoudig dan een rechte lijn.

Interstitiële longziekten

De longen zorgen voor gasuitwisseling tussen het bloed en de lucht. Zuurstof gaat van de lucht naar het bloed en koolstofdioxide van het bloed naar de lucht. Figuur 2 toont een schematische tekening van de longen (links). Ingeademde lucht stroomt door de luchtpijp via de bronchiën naar de bronchiolen en tenslotte naar de longblaasjes (rechtsboven), waar de gasuitwisseling plaatsvindt. Het pulmonale interstitium is een netwerk van weefsel ter ondersteuning van de longblaasjes (linksonder). Interstitiële longziekten (ILD) is de verzamelnaam voor een diverse groep van meer dan 200 ziekten die het interstitium in de longen aantasten. Prognose en behandeling zijn afhankelijk van de exacte aard van de specifieke ILD en daarom is het stellen van de juiste diagnose van cruciaal belang. Het stellen van die diagnose vereist een interdisciplinaire benadering, waarbij longfunctietesten, medische beeldvorming en laboratoriumtests een rol spelen. CT-scans zijn hierin een belangrijke factor.

Doel van dit proefschrift

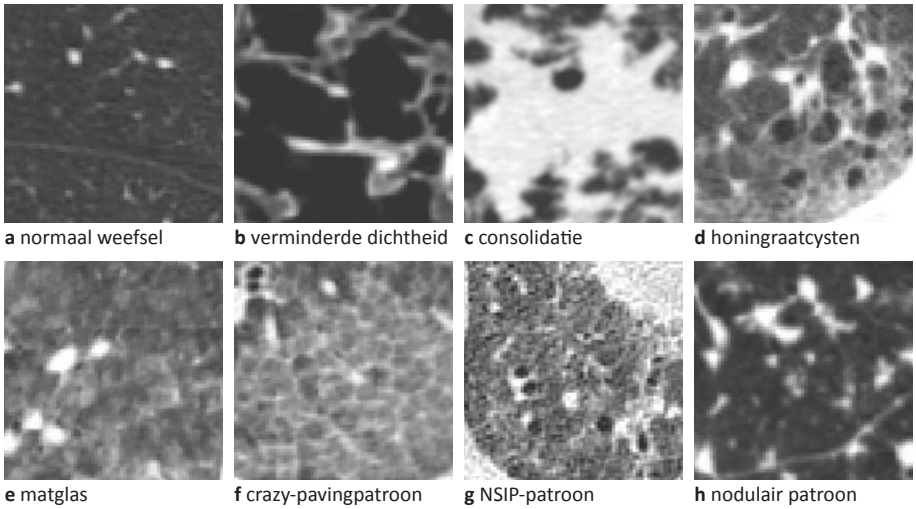
Voor analyse van CT scans van de longen is het vaak nodig om in de scan bepaalde afwijkingen of structuren in te tekenen, of te annoteren. Deze annotaties kunnen bijvoorbeeld gebruikt worden voor het stellen van een diagnose of voor het bepalen van het effect van een behandeling. Voor sommige vraagstukken kan dat kan automatisch met behulp van een classificatiealgoritme, maar er zijn ook problemen waarbij automatische classificatie niet goed genoeg werkt. In dat geval kan een expert de scan handmatig annoteren, met als nadeel dat dit erg arbeidsintensief is. Het doel van het werk beschreven in dit proefschrift is het vergemakkelijken van het annoteren van structuren en afwijkingen in CT-scans van de borstkas. We ontwikkelden een interactief computerprogramma, dat door een expert getraind kan worden om verschillende texturen onderscheiden. Voorbeelden van deze texturen zijn 'normaal longweefsel' of 'longweefsel met een verminderde dichtheid'. De interactieve methode hebben we toegepast op verschillende vraagstukken: longsegmentatie in CT-scans van dieren en mensen met ernstige longafwijkingen, classificatie van normaal en verschillende soorten abnormaal longweefsel bij patiënten ILD en classificatie van normaal en aangetast longweefsel bij patiënten op de intensive care (IC).



Figuur 3. Stroomdiagram van interactieve annotatie. De basis van het proces is de annotatiecyclus, waarin een expert de resultaten van automatische classificatie van VOI's in een transversale coupe corrigeert, waarna deze VOI's worden toegevoegd aan de verzameling voorbeelden waarmee het classificatiealgoritme wordt getraind. Het algoritme leert om de verschillende texturen die te zien zijn in de scan te onderscheiden door middel van de correcties van de expert. Als een vooraf bepaald percentage van de scan is geannoteerd is, wordt het classificatiealgoritme een laatste keer getraind en classificeert het alle overgebleven VOI's. De expert verbetert daarna de laatste fouten.

De eerste stap bij interactieve annotatie is het automatisch afbakenen of segmenteren van de structuur waarin de expert, een arts of onderzoeker, geïnteresseerd is. Dit kunnen bijvoorbeeld de borstkas of de longen zijn. Daarna wordt deze structuur onderverdeeld in kleinere delen (volumes of interest, of VOI's), die zo gemaakt zijn dat er maar één textuur in voorkomt. Deze VOI's krijgen vervolgens interactief een label.

Figuur 3 toont een stroomschema van de interactieve methode voor het annoteren van structuren in CT-scans. Annotatie vindt plaats in cycli, waarbij een classificatiealgoritme iedere keer de labels van de VOI's in een deel van de scan voorspelt en een expert de labels van verkeerd geclassificeerde VOI's verbetert. Het classificatiealgoritme leert van de verbeteringen en wordt zodoende steeds beter in het herkennen van de verschillende texturen in de scan, hierbij rekening houdend met de manier van annoteren van de expert. Dit proces van correctie, training en classificatie gaat door tot een vooraf bepaald percentage van het longweefsel is beoordeeld, of tot de expert vindt dat het classificatiealgoritme optimaal getraind is. Dan wordt het algoritme een laatste keer getraind, waarna het de resterende VOI's classificeert. Tenslotte loopt de expert de hele scan nog een keer na om de laatste fouten te verbeteren.



Figuur 4. Voorbeelden van normaal (**a**) en verschillende typen abnormaal (**b-h**) longweefsel.

Als alternatief kan de expert ervoor kiezen om alle VOI's handmatig te labelen, of om automatische classificatieresultaten van alle VOI's in één keer te corrigeren, zonder het classificatiealgoritme tussendoor te gebruiken.

Hoofdstuk 2 beschrijft de toepassing van interactieve annotatie op CT-scans van de longen van patiënten met ILD. Het doel was hierbij om een instrument te ontwikkelen voor het annoteren van normaal en zeven soorten abnormaal longweefsel, namelijk verminderde dichtheid, consolidatie, honingraatcysten, matglas, crazy-pavingpatroon, NSIP-patroon en nodulair patroon (figuur 4). In dit hoofdstuk werden computersimulaties van het gedrag van een expert gebruikt om verschillende interactieve annotatiestrategieën te vergelijken. Bij de eerste strategie classificeerde een algoritme alle VOI's in een scan automatisch. Het algoritme was getraind met VOI's uit andere, eerder geannoteerde scans. In dit scenario moest de expert het label van gemiddeld 58% van alle VOI's veranderen. De tweede methode maakte gebruik van de interactieve annotatiemethode uit figuur 3. De eerste coupe werd gelabeld met behulp van een heuristische benadering: alle VOI's kregen het label 'normaal weefsel', omdat dat deze categorie het meest voorkwam in de gebruikte dataset. De VOI's in de volgende coupes werden interactief geannoteerd. Bij deze aanpak moest de expert het label van gemiddeld 21% van alle VOI's corrigeren. De derde methode was gelijk aan de tweede, met uitzondering van de labeling van de VOI's in de eerste coupe. Deze stap werd uitgevoerd door een classificatiealgoritme getraind met behulp van VOI's afkomstig van scans die al eerder door een expert waren geannoteerd. Gemiddeld moest de expert bij deze laatste methode het label van 20% van de VOI's aanpassen.

Hoofdstuk 3 onderzocht optimalisatiestrategieën voor interactieve annotatie uit hoofdstuk 2, met als doel het aantal incorrect gelabelde VOI's en daarmee de hoeveelheid werk voor de expert te minimaliseren. Hierbij werden dezelfde texturen als in hoofdstuk 2 onderscheiden: normaal weefsel, verminderde dichtheid, consolidatie, honingraatcysten, matglas, crazy-pavingpatroon, NSIP-patroon en nodulair patroon. Met behulp van simulatiesoftware hebben we verschillende scenario's getest. Eerst werden experimenten met automatische verificatiemethoden uitgevoerd. De prestaties van vier classificatiealgoritmes werden vergeleken: een algoritme getraind met de annotaties van één expert, een algoritme getraind met annotaties van drie experts, een algoritme getraind met annotaties waarover alle experts het eens waren en een verzameling van classificatiealgoritmes. In dat laatste geval werd elk classificatiealgoritme in de verzameling getraind met data afkomstig van een andere bron. Het label dat door de meeste classificatiealgoritmes in de verzameling werd voorspeld, was het label dat werd toegekend. De experimenten met de vier bovengenoemde methoden werden allemaal twee keer uitgevoerd, eenmaal met en eenmaal zonder textuurselectie. Zonder textuurselectie werden VOI's van alle acht texturen gebruikt om het classificatiealgoritme te trainen. Met textuurselectie werden alleen VOI's van de textuurklassen die aanwezig waren in de scan gebruikt. In een tweede reeks experimenten werd interactieve annotatie gesimuleerd om de effecten van de volgende aanpassingen op het standaard algoritme voor interactieve annotatie te evalueren: (1) textuurselectie, (2) het gebruik van een classificatiealgoritme getraind op eerdere annotaties en (3) het aanbieden van een keuze tussen verschillende interactieve of automatische classificatieresultaten als uitgangspunt voor de correctie van labels in een coupe. Voor alle experimenten werd berekend hoe goed de labels van VOI's voorspeld konden worden. Het best presterende protocol was dat waarin experts textuurselectie toepasten en waarin ze konden kiezen welke classificatieresultaten ze wilden gebruiken als uitgangspunt voor correctie. Gemiddeld werd bij deze strategie het label van 88% van de VOI's correct voorspeld. In de laatste reeks experimenten werden strategieën voor de selectie van de coupes die werden voorgelegd aan de expert geëvalueerd. De beste methode bleek die waarbij te expert de coupes te zien kreeg waarover het classificatiealgoritme het minst zeker was. Het maakte daarbij niet uit of de coupes steeds uit een ander deel van de scan gekozen moesten worden of niet.

Hoofdstuk 4 beschrijft het gebuik van interactieve annotatie voor longsegmentatie. Tweeëndertig CT-scans werden hiervoor gebuikt: acht scans van longtransplantatiepatiënten, acht scans uit de LOLA11 challenge, acht scans van varkens en acht micro-CT-scans van muizen. In alle scans werd de borstkas automatisch gesegmenteerd en verdeeld in VOI's. VOI's met een gemiddelde waarde van -500 HU of minder werden automatisch gelabeld als longweefsel, de rest als niet-longweefsel. Een expert verfijnde deze segmenta-

tie interactief, hetzij door het steeds opnieuw trainen van een classificatiealgoritme, hetzij door het één-voor-één corrigeren van automatisch gegenereerde labels. De resultaten hiervan werden vergeleken met handmatige segmentatie van de longen. Met de eerste aanpak duurde interactieve longsegmentatie gemiddeld 9 minuten. Hierbij moest de expert gemiddeld 2,0% van alle labels in een scan aanpassen. Met behulp van de tweede methode duurde longsegmentatie gemiddeld 13 minuten, waarbij de expert gemiddeld de labels van 3,0% van alle VOI's in een scan aan moest passen. De interactief verkregen segmentaties kwamen goed overeen met handmatige segmentaties, zo bleek bij evaluatie van acht coupes per scan. De gemiddelde Dice-coëfficiënt was 0,933.

Tot slot behandelde **hoofdstuk 5** de toepassing van interactieve longsegmentatie gevolgd door textuuranalyse in 30 CT-scans van patiënten die beademd werden op de intensive care. Beademing is een potentieel levensreddende behandeling, maar kan leiden tot verdere schade aan de longen bij patiënten met acute respiratory distress syndrome (ARDS) wanneer te grote ademvolumes worden toegepast. In de klinische praktijk wordt het ademvolume bepaald op basis van lichaamslengte van de patiënt en op basis van de aanwezigheid of afwezigheid van ARDS. Deze benadering houdt echter geen rekening met de hoeveelheid longweefsel dat nog normaal belucht is. Hoe kleiner deze hoeveelheid, hoe kleiner het ademvolume dat veilig kan worden toegepast. Het doel van dit hoofdstuk was om met behulp van CT-scans de totale longcapaciteit en het volume van normaal en deels belucht longweefsel van patiënten te bepalen. Longen werden interactief gesegmenteerd met behulp van de in hoofdstuk 4 beschreven methode en vervolgens verdeeld in kleinere VOI's. Deze VOI's kregen een een van de volgende labels: hyperinflatie, normaal belucht, slecht belucht en niet-belucht, op basis van hun gemiddelde CT waarde. We vonden dat de totale longcapaciteit en het normaal beluchte longvolume niet betrouwbaar konden worden geschat op basis van alleen lichaamslengte en de aanwezigheid of afwezigheid van ARDS. Met behulp van de beschreven interactieve benadering kunnen betrouwbare schattingen worden gemaakt voor individuele patiënten. Dit is een eerste stap naar gepersonaliseerde berekening van het optimale ademvolume.

List of publications

Papers in international journals

TTJP Kockelkorn, PA de Jong, CM Schaefer-Prokop, R Wittenberg, AM Tiehuis, HA Gietema, JC Grutters, MA Viergever, B van Ginneken. Semi-automatic classification of textures in thoracic CT scans. *Physics in Medicine and Biology*. **2016**;61(16):5906-24

J Ramos, **TTJP Kockelkorn**, I Ramos, R Ramos, B van Ginneken, MA Viergever, A Campilho, JC Grutters. Content-Based Image Retrieval by Metric Learning from Radiology Reports: Application to Interstitial Lung Diseases. *IEEE Journal of Biomedical and Health Informatics.* **2016**;1(20):281-292.

TTJP Kockelkorn, CM Schaefer-Prokop, G Bozovic, A Muñoz-Barrutia, EM van Rikxoort, MS Brown, PA de Jong, MA Viergever, B van Ginneken. Interactive lung segmentation in abnormal human and animal chest CT scans. *Medical Physics.* **2014**;8(41):081915.

M Arai, N Obata, **TTJP Kockelkorn**, K Yamada, T Toyota, S Haga, Y Yoshida, H Ujike, I Sora, I Ikeda, T Yoshikawa, M Itokawa. Lack of association between polymorphisms in the 5' upstream region of the DISC1 gene and mood disorders. *Psychiatr Genet.* **2007**;6(17):357.

P Kemmeren, **TTJP Kockelkorn**, T Bijma, R Donders, FC Holstege, Predicting gene function through systematic analysis and quality assessment of high-throughput data. *Bioinformatics.* **2005**;8(21):1644-1652.

M Radonjic, JC Andrau, P Lijnzaad, P Kemmeren, **TTJP Kockelkorn**, D van Leenen, NL van Berkum, FC Holstege. Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S. cerevisiae* stationary phase exit. *Mol. Cell.* **2005**;2(18):171-183.

J van de Peppel, N Kettelarij, H van Bakel, **TTJP Kockelkorn**, D van Leenen, FC Holstege. Mediator expression profiling epistasis reveals a signal transduction pathway with antagonistic submodules and highly specific downstream targets. *Mol. Cell.* **2005**;4(19):511-522.

TTJP Kockelkorn, M Arai, H Matsumoto, N Fukuda, K Yamada, Y Minabe, T Toyota, H Ujike, I Sora, N Mori, T Yoshikawa, M Itokawa. Association study of polymorphisms in the 5' upstream region of human DISC1 gene with schizophrenia. *Neurosci Lett.* **2004**;1(368):41-45.

Papers in conference proceedings

TTJP Kockelkorn, R Ramos, J Ramos, CI Sánchez, PA de Jong, CM Schaefer-Prokop, JC Grutters, B van Ginneken, MA Viergever. Interactive classification of lung tissue in CT scans by combining prior and interactively obtained training data: a simulation study. In: *International Conference on Pattern Recognition.* **2012**:105-108

J Ramos, **TTJP Kockelkorn**, B van Ginneken, MA Viergever, R Ramos, A Campilho. Supervised Content Based Image Retrieval Using Radiology Reports. In: *Image Analysis and Recognition.* **2012;**7325:249-258.

TTJP Kockelkorn, PA de Jong, HA Gietema, JC Grutters, M Prokop, B van Ginneken. Interactive annotation of textures in thoracic CT scans. In: *SPIE Medical Imaging.* **2010**; 7624:76240X.

TTJP Kockelkorn, EM van Rikxoort, JC Grutters, B van Ginneken. Interactive lung segmentation in CT scans with severe abnormalities. In: *IEEE International Symposium on Biomedical Imaging.* **2010**:564-567.

EM van Rikxoort, M Galperin-Aizenberg, JG Goldin, **TTJP Kockelkorn**, B van Ginneken, MS Brown. Multi-classifier semi-supervised classification of tuberculosis patterns on chest CT scans. In: *The Third International Workshop on Pulmonary Image Analysis.* **2010**:41-48.

CI Sánchez, M Niemeijer, **TTJP Kockelkorn**, MD Abramoff, B van Ginneken. Active Learning Approach for Detection of Hard Exudates, Cotton Wool Spots and Drusen in Retinal Images. In: *SPIE Medical Imaging.* **2009**;7260:72601I.

Interactief dankwoord

Inleiding

Het dankwoord is een belangrijk, maar zeker niet triviaal onderdeel van een proefschrift. Zeker onder tijdsdruk kan het gebeuren dat promovendi mensen vergeten te bedanken, hetgeen tot vervelende situaties kan leiden. Het vermoeden bestaat dat naarmate de promotieperiode langer duurt, de kans op omissies in het dankwoord toeneemt (ongepubliceerde gegevens). Gezien de hoge impact enerzijds en de hoge foutgevoeligheid anderzijds, stel ik een interactief algoritme voor, waarbij de lezer van het dankwoord middels een beslisboom erachter komt wie waarvoor bedankt wordt.

Materialen en methoden

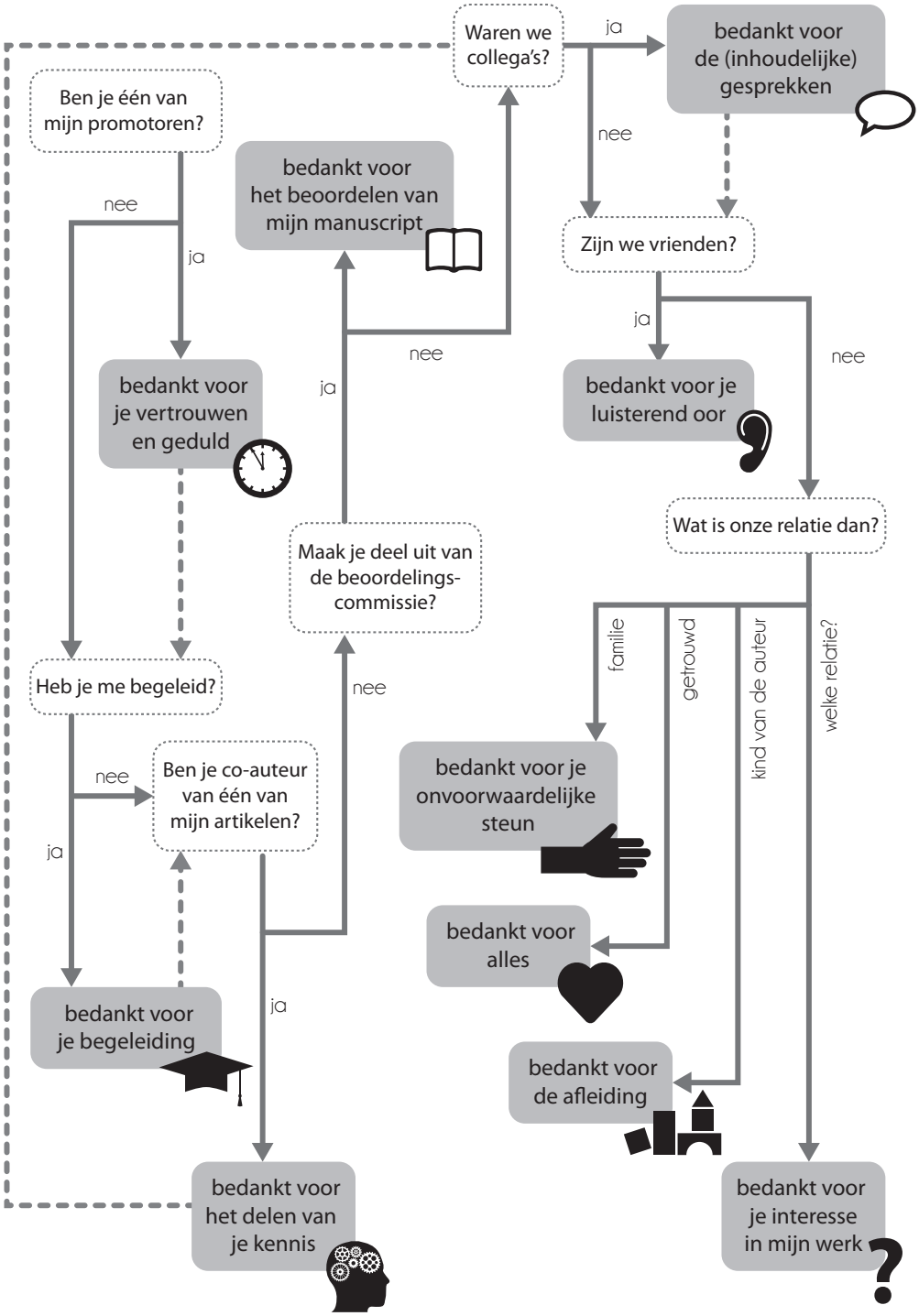
De beslisboom werd ontwikkeld in Adobe Illustrator CC (release 2015.3), op basis van persoonlijke observaties in de periode 2008-2016. Lezers dienen te beginnen bij de vraag "Ben je één van mijn promotoren?", linksboven in het diagram. De mogelijke antwoorden op elke vraag leiden naar nieuwe vakken met vragen, totdat er een grijs veld wordt bereikt, met daarin woorden van dank. Uiteraard kunnen mensen ook op meerdere gebieden bedankt worden; daarom leiden er routes vanuit sommige dankvelden naar nieuwe vragen.

Resultaten

Figuur 1 geeft de beslisboom weer. Omdat de het algoritme nog niet gevalideerd is, wil ik een aantal mensen ook nog op meer traditionele wijze bedanken, te beginnen met mijn promotoren, Max Viergever, Bram van Ginneken en Mathias Prokop.

Bram, dank voor de vrijheid die je me hebt gegeven om aan de stukken in dit proefschrift te werken. Je directe manier van feedback geven is de kwaliteit van de artikelen zeker ten goede gekomen! Beste Max, dank voor je begeleiding in de afgelopen jaren, maar vooral voor je vertrouwen in een goede afloop. Dat vertrouwen was van doorslaggevend belang voor het afkomen van mijn proefschrift. Mathias, je enthousiasme voor dit vakgebied werkt aanstekelijk. Dank voor je inspiratie tijdens de Pulmo meetings.

Daarnaast hebben Pim de Jong en Cornelia Schaefer-Prokop een belangrijke rol gespeeld in het tot stand komen van dit werk. Beste Pim, of het nu om het maken van annotaties ging of om het nalezen van een manuscript, je antwoord kwam altijd razendsnel. Dank voor je positieve en oplossingsgerichte inbreng. Cornelia, ik heb op klinisch gebied veel van je geleerd. Dank voor je hulp bij het vormgeven en uitvoeren van de studies.



Figuur 1. Interactieve beslisboom om te bepalen wie waarvoor bedankt wordt.

Dear co-authors, Gracijela Bozovic, Arrate Muñoz-Barrutia, Eva van Rikxoort, Matthew Brown, Rianne Wittenberg, Audrey Tiehuis, Hester Gietema, Jan Grutters, Rui Ramos, José Ramos, Dave Dongelmans, and Ludo Beenen, thank you all for your valuable contributions to chapters 2, 3, 4, and 5.

Beste leden van de beoordelingscommissie, prof. dr. Lammers, prof. dr. Leiner, prof. dr. Raaymakers, prof. dr. ir. Karssemeijer en dr. de Bruijne, hartelijk dank voor het beoordelen (en goedkeuren) van mijn proefschrift.

Beste ISI-collega's, zonder jullie was het schrijven van dit proefschrift vast veel moeizamer verlopen. Bedankt voor de inhoudelijke discussies, praktische hulp en luisterende oren. Een speciaal woord van dank aan Gerard voor het reanimeren van mijn computer. Chris, dank voor het meermaals bemiddelen bij de ruzies tussen mij en C++. Hugo, Renée, Jacqueline en Marjan, zeker in de laatste periode heb ik veel steun aan jullie gehad. Heel veel dank daarvoor! Adriënne en Anneriet, mijn paranimfen, bedankt dat ik altijd bij jullie terecht kon. Wat fijn dat jullie ook tijdens de verdediging achter me staan.

Lieve vrienden, familie en schoonfamilie, dank voor jullie blijvende interesse in mijn onderzoek, maar vooral ook voor het zo nu en dan tactisch laten rusten van het onderwerp. Lieve omi, papa, mama en Ward, jullie zijn er altijd voor me. Dank voor jullie onvoorwaardelijke steun.

Lieve Job, Sterre, Pepijn en Casper, vier betere redenen om zo lang over mijn promotie te doen kan ik me niet voorstellen! En tenslotte, lieve Allard, in de hectiek van de afgelopen jaren ben je altijd mijn houvast geweest. Samen kunnen we de wereld aan!

Conclusie

Dit proefschrift is tot stand gekomen met de hulp van velen. Dank voor jullie bijdragen, groot of klein, inhoudelijk of persoonlijk!

Curriculum vitae

Thessa Kockelkorn (15 December 1978, Heerlen) obtained her BSc summa cum laude at University College Utrecht in 2000. She studied Biomedical Sciences, specializing in Bioinformatics, at Utrecht University and received her MSc cum laude in 2002. Thessa worked as a researcher at the Tokyo Institute of Psychiatry and at the University Medical Center Utrecht, after which she transferred to industry. She worked as an independent graphic and web designer, and as a business analyst at Plexus Medical Group (now KPMG Health). In May 2008, Thessa returned to academia for her PhD research at the Image Sciences Institute, focusing on interactive texture analysis in chest CT scans. The main results of her work are described in this thesis.

As of September 2016, Thessa works as a data scientist at EdgeLeap B.V. in Utrecht. Thessa is married to Allard and together they have four children: Job (2007), Sterre (2008), Pepijn (2011), and Casper (2014).