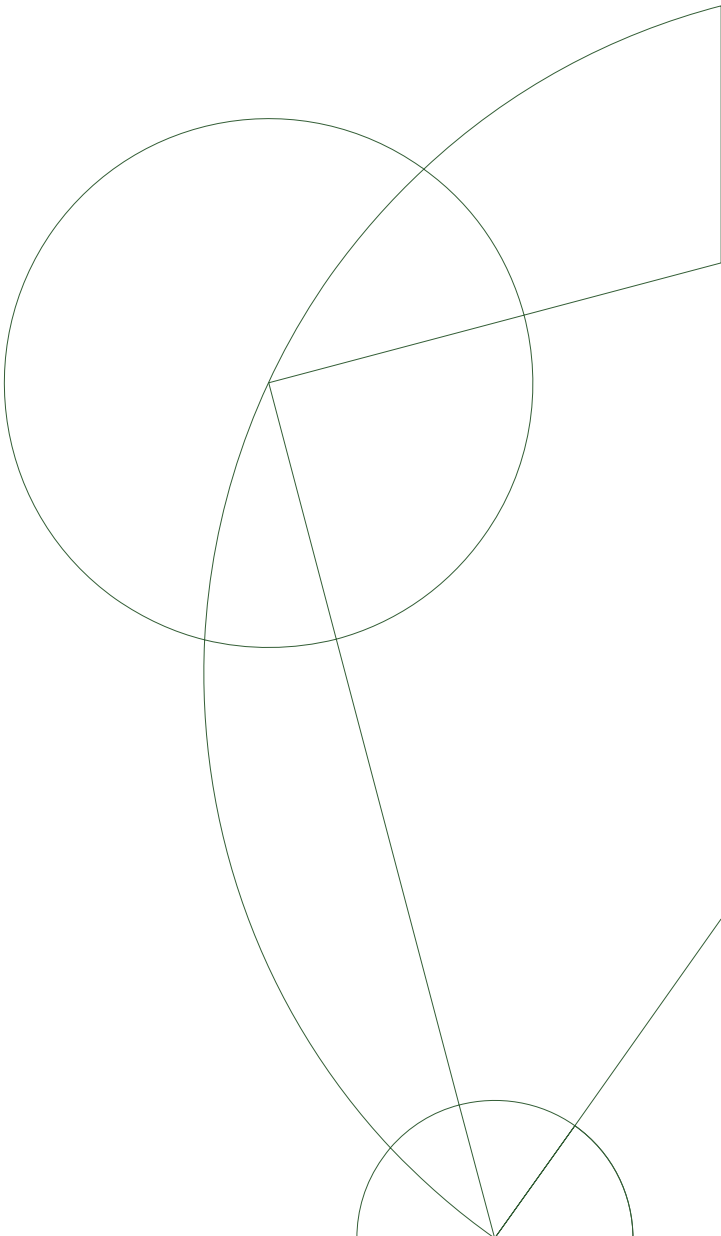
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

PhD thesis

Lauge Sørensen

Pattern Recognition-Based Analysis of COPD in CT



Academic advisor: Marleen de Bruijne

Co-supervisors: Mads Nielsen and Jon Sporring

Submitted: 02/06/10

Abstract

Computed tomography (CT), a medical imaging technique, offers a detailed view of the human body that can be used for direct inspection of the lung tissue. This allows for in vivo measurement of subtle disease patterns such as the patterns associated with chronic obstructive pulmonary disease (COPD). COPD, also commonly referred to as "smokers' lungs", is a lung disease characterized by limitation of the airflow to and from the lungs causing shortness of breath. The disease is expected to rank as the fifth most burdening disease worldwide by 2020 according the the World Health Organization. COPD comprises two main components, chronic bronchitis, characterized by inflammation in the airways, and emphysema, characterized by loss of lung tissue. Emphysema basically looks like black blobs of varying sizes within the normal, gray lung tissue in CT, and can therefore be seen as a family of texture patterns. Commonly employed CT-based quantitative measures in the clinical literature are rather simplistic and do not take the texture appearance of the lung tissue into account. This includes measures such as the relative area (RA), also called emphysema index, that applies a fixed threshold to each individual lung voxel in the CT image and counts the number of voxels below the threshold relative to the total amount of lung voxels.

This thesis presents several methods for texture-based quantification of emphysema and/or COPD in CT images of the lungs. The methods rely on image processing and pattern recognition. The image processing part deals with characterizing the lung tissue texture using a suitable texture descriptor. Two types of descriptors are considered, the local binary pattern histogram and histograms of filter responses from a multi-scale Gaussian derivative filter bank. The pattern recognition part is used to turn the texture measures, measured in a CT image of the lungs, into a quantitative measure of disease. This is done by applying a classifier that is trained on a training set of data examples with known lung tissue patterns. Different classification systems are considered, and we will in particular use the pattern recognition concepts of *supervised learning*, *multiple instance learning*, and *dissimilarity representation-based classification*.

The proposed texture-based measures are applied to CT data from two different sources, one comprising low dose CT slices from subjects with manually annotated regions of emphysema and healthy tissue, and one comprising volumetric low dose CT images from subjects that are either healthy or suffer from COPD. Several experiments demonstrate that it is clearly beneficial to take the lung tissue texture into account when classifying or quantifying emphysema and/or COPD in CT. Compared to RA and other common clinical CT-based measures, the texture-based measures are better at discriminating between CT images from healthy and COPD subjects, they correlate better with the lung function of the subjects, they are more reproducible, and they are less influence by the inspiration level of the subject during CT scanning – a major source of variability in CT.

Contents

Chapter 1

Introduction

This thesis presents several learning-based methods for quantitative analysis of emphysema and/or chronic obstructive pulmonary disease (COPD) in computed tomography (CT) images of the lungs. The methods basically rely on pattern recognition techniques for classifying lung tissue as described using texture descriptors in order to arrive at a measure of disease. *Learning-based* refers to the fact that a classifier is trained on data represented by a suitable set of texture descriptors that is learned from data examples. The work is close to the field of computer-aided detection/diagnosis (CAD). Although, the definition of CAD varies.

The general picture seems to be that the role of CAD is to provide a tool that human experts can use to aid in arriving at a final diagnostic decision. However, it is still the human that has the final say. This view is reflected in several recent CAD review papers:

> *"With CAD, radiologists use the computer output as a "second opinion," and radiologists make the final decisions."* [23]
>
> *"If successfully developed, CAD can be a useful second opinion to radiologists in thoracic CT interpretation."* [13]
>
> *"Radiologists were expected to ultimately use the output from computerized analysis of medical images as a "second opinion," like a spellchecker, in detecting and characterizing lesions as well as in making diagnostic decisions."* [34]

When the application is diagnosis of individuals and false negatives have large consequences, e.g., missing a malignant lung nodule, CAD as a second opinion for a human expert is the way to go. However, when a whole population is studied for general trends, e.g., in epidemiology, a fully-automated approach is preferable. Further, since more and more data is being produced, both in daily clinical practice and in screening trials, and there is a need for analyzing all this data, fully-automated approaches become increasingly important.

Historically, CAD often focuses on detection and/or diagnosis of focal abnormalities, e.g., lung nodules [13, 23, 86, 104], vertebral fractures [23], aneurysms [23], and pulmonary embolisms [13,86]. Although, detection of breast cancer in mammography, where signs are more subtle, also has received a lot of attention [34].

The scope of this thesis is quantification of subtle disease patterns associated with COPD, mainly emphysema but also chronic bronchitis, covering larger regions in the lung. We also aim at proceeding beyond current knowledge of human experts, by proposing methods that do not require human experts to annotate training data. Few published studies working on classification of abnormal lung tissue in CT using texture information proceed beyond the classification of individual regions of interest (ROIs) [86]. To achieve the goal of quantitatively measure disease in subjects based on medical images, the images need to be analyzed on a global scale and information from the entire images should be combined. This is effectively what we do in this thesis, with focus on COPD, namely, investigate learning-based methods that outputs a single measure for an entire, possibly three-dimensional, CT image based on combining information from several ROIs.

1.1 Chronic obstructive pulmonary disease (COPD)

Chronic obstructive pulmonary disease (COPD) is an umbrella term covering several diseases. A general definition is quoted below:

> *"Chronic obstructive pulmonary disease (COPD) is characterized by the progressive development of airflow limitation that is not fully reversible. The term COPD encompasses chronic obstructive bronchitis, with obstruction of small airways, and emphysema, with enlargement of air spaces and destruction of lung parenchyma, loss of lung elasticity, and closure of small airways."* [7]

The relative contribution of the two main components, emphysema and chronic bronchitis, vary from person to person [7], and there is a complex interrelationship among these components that results in the progressive reduction in expiratory airflow [73]. COPD is a major public health problem. It is the fourth leading cause of morbidity and mortality in the United States alone, and is predicted to rise from its ranking in 2000 as the 12th most prevalent disease worldwide to the 5th, and from the 6th most common cause of death to the 3rd by 2020 [7,77]. The dramatic increase in COPD is amongst others because of reduced mortality from other causes, such as cardiovascular diseases, and a marked increase in cigarette smoking [7].

Tobacco smoking is the most important and well-studied risk factor in developing COPD [7,40,52,73,77]. Although, other less-studied factors, such as genes and indoor and outdoor air pollution, also play a role [7, 77]. A recent study showed that after 25 years of smoking, at least 25% of smokers without initial disease had clinically significant COPD and $30 - 40\%$ had any type of COPD [52]. The current gold standard for measuring the airflow limitation associated with COPD is by means of

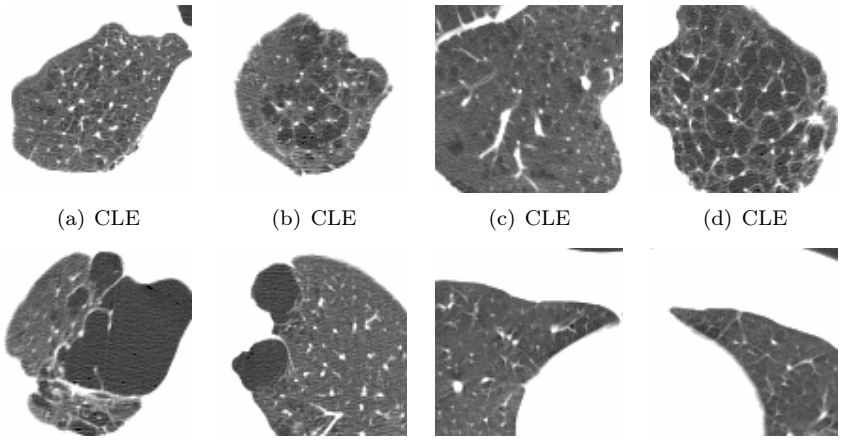
Table 1.1: Spirometric classification of COPD severity [77]. If the condition $FEV_1/FVC < 0.7$ holds, the subject is diagnosed with COPD, and the severity is determined according to the conditions given in the table, otherwise the classification is no COPD.

GOLD stage	condition
stage I (mild)	$FEV_1\%pred \geq 80\%$
stage II (moderate)	$50\% \leq FEV_1\%pred < 80\%$
stage III (severe)	$30\% \leq FEV_1\%pred < 50\%$
stage IV (very severe)	$FEV_1\%pred < 30\%$

spirometry [73,77], the most common pulmonary function test (PFT), and the Global Initiative for Chronic Obstructive Lung Disease (GOLD) [77] has defined the so-called GOLD stages; a simple spirometric classification of COPD severity into four stages. This is the objective ground truth that is used in this thesis. The subject breathes into a mouthpiece that is connected to an instrument called a spirometer, and specific measurements are taken during this procedure. Measures important for the diagnosis of COPD are: the forced vital capacity (FVC), defined as the volume of air, measured in liters, that one can forcibly blow out after full inspiration; and forced expiratory volume in one second (FEV_1), defined as the maximum volume of air, measured in liters, that one can forcibly blow out in the first second during the FVC manoeuvre. The diagnosis according to the GOLD stages is listed in Table 1.1. The condition $FEV_1/FVC < 0.7$ confirms the presence of airflow limitation, and thereby COPD, since the subject is able to exhale less than 30% of the total capacity during the first second of exhalation. The severity is determined according to how much air is actually exhaled during the first second, as measured by FEV_1 corrected according to reference values based on age, height, sex, and race. This value is termed $FEV_1\%pred$ in this thesis.

Unfortunately, PFTs are insensitive to early stages of COPD, approximately 30% of the lung must be destroyed by emphysema before being detectable by PFTs [40], for example. CT, an X-ray-based digital imaging modality capable of visualizing the inside of the human body, has emerged as an alternative method that can facilitate direct, in vivo measurement of the components of COPD. The values in a CT image, called Hounsfield units (HU), are measures of the radiation attenuation coefficient of the tissue displayed with respect to the radiation attenuation coefficient of water. Two points are fixed on the HU scale, water is defined as 0 HU and air as -1000 HU, and the HUs in a CT image can therefore be directly related to the density of the displayed tissue, provided that the scanner is calibrated properly.

Emphysema lesions are visible in CT images as areas of abnormally low attenuation values close to that of air, i.e., black "blobs" within the gray lung tissue. In CT, emphysema can be classified into three subtypes, or patterns, and we will adopt the naming and definitions used in Webb *et al.* [108]. These subtypes are the following: centrilobular emphysema (CLE), defined as multiple small low-attenuation areas; paraseptal emphysema (PSE), defined as multiple low-attenuation areas in a single



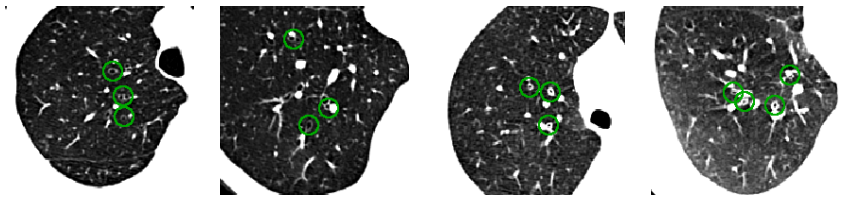
(a) CLE (b) CLE (c) CLE (d) CLE

(e) PSE (f) PSE (g) PSE (h) PSE

Figure 1.1: Examples of emphysema patterns of varying severity. The examples are from CT slices where the leading emphysema pattern was determined by consensus of an experienced chest radiologist and a CT experienced pulmonologist. All examples are shown with the recommended window setting of -700/1000 HU [108], where the two values are: [the center of the window]/[the width of the window]. White pixels are densities above 200 HU, including the exterior of the lung, the vessels, and the airway walls. Black is missing lung tissue due to emphysema, except for the top-right part of (a) which is part of the trachea, and the top-right part (g) and (h) which is outside the body. Gray is lung tissue.

layer along the pleura often surrounded by interlobular septa that is visible as thin white walls; and panlobular emphysema (PLE), defined as a low-attenuation lung with fewer and smaller pulmonary vessels. CLE and PSE mostly appear in smokers, whereas PLE is associated with a genetic disease [83]. We will therefore mainly encounter CLE and PSE since the CT images used in this thesis are from current or former smokers that do not suffer from the aforementioned genetic disease. Examples of CLE and PSE are shown in Figure 1.1. The pathological changes in chronic bronchitis include obstruction of the small airways due to inflammation and fibrosis [83]. The airways are generally visible in CT as bright "tubes" of a certain thickness, the airway walls, with a dark interior, the lumen. Inflammation and fibrosis causes the airway walls to become thicker. Examples of normal airways and airways with suspected chronic bronchitis are marked in green in Figure 1.2. The marked airways are perpendicular to the image plane.

Visual and computerized assessment in CT images has emerged as an alternative to PFTs that directly can measure the two components of COPD. However, it is difficult to visually assess disease severity and progression. A human is good at recognizing and distinguishing texture patterns, but poor at precisely determining the fraction of the lung that is represented by a particular texture pattern. Moreover, visual assessment is subjective, time-consuming, and suffers from intra-observer



(a) No inflammation (b) No inflammation (c) Inflammation (d) Inflammation

Figure 1.2: CT slices with examples of no inflammation/inflammation in the airways marked in green. The airway can be seen as a black spot, the lumen, surrounded by a white wall, the airway wall. The inflammation is visible as thick airway walls and, as a consequence of this, a narrowed airway.

(a) Original axial CT slice (b) RA_{950} (shown in black) (c) RA_{910} (shown in black)

Figure 1.3: An axial slice from a subject in the DLCST database with severe COPD according to spirometry: $FEV_1/FVC = 0.37$ and $FEV_1\%pred = 0.32$. Also the extent of emphysema is vast according to RA, 51.3% with a threshold of -950 HU. Note that RA counts all that is shown in black in (b) and (c) except for the trachea that is visible as a black region in between the two lungs.

and inter-observer variability [6, 58]. Computerized assessment does not suffer from these limitations, and the sensitivity to emphysema and/or COPD, as well as the reproducibility, of lung density parameters computed from CT images are superior to PFTs [83]. A widely used density parameter computed from CT is the relative area of emphysema (RA), also referred to as emphysema index [83]. RA is basically a thresholding technique. The lung tissue is thresholed according to a certain HU threshold close to -1000 HU, i.e., close to the density of air, and the area of voxels below the threshold relative to the toal amount of lung tissue voxels is reported. An example axial CT slice from a subject with severe COPD is shown in Figure 1.3 together with thresholded versions of the slice obtained by applying thresholds of -910 and -950 HU, respectively. Clearly, RA disregards a lot of the information that is available in CT. For example, by summarizing the entire distribution of attenuation values within the lungs by a single measure, the percentage of voxels below the HU threshold, but also by restricting the analysis to a single threshold and by considering the intensity in each voxel independently. The aim of this thesis is to develop CT-based measures of COPD that take more information into account, thereby improving upon the existing simple lung density measures such as RA.

Two sources of data are used in this thesis. The first source is a data set from an exploratory study carried out at the Department of Respiratory Medicine at Gentofte University Hospital Denmark comprising low dose CT slices obtained at the upper, middle, and lower part of the lungs from 40 subjects (20 with moderate to severe COPD according to PFTs, 10 asymptomatic smokers, and 10 healthy non-smoking volunteers) [84]. The second source is the Danish Lung Cancer Screening Trial (DLCST) [67], a screening trial where more than 2000 former and current smokers with a smoking history of more than 20 pack years[1] were scanned annually for five consecutive years. These are low dose three-dimensional CT images that consist of a stack of approximately 400 CT slices each. An example of a CT image from the DLCST database is shown in Figure 1.4. The three planes shown are the commonly used views, and they consist of the conventional axial plane, which is also the plane use in the data set from [84], and the two orthogonal planes to the axial plane.

1.2 Pattern recognition, and image processing, for analysis of lung disease

A text book example [24] of how the structure of many systems for pattern recognition can be viewed is provided in the following. From input data, a decision is made by propagating the input through the following steps in the order given:

Sensing images or sounds or other physical input is converted into signal data.

Segmentation sensed objects are isolated from the background or from other objects.

Feature extraction object properties useful for classification are measured.

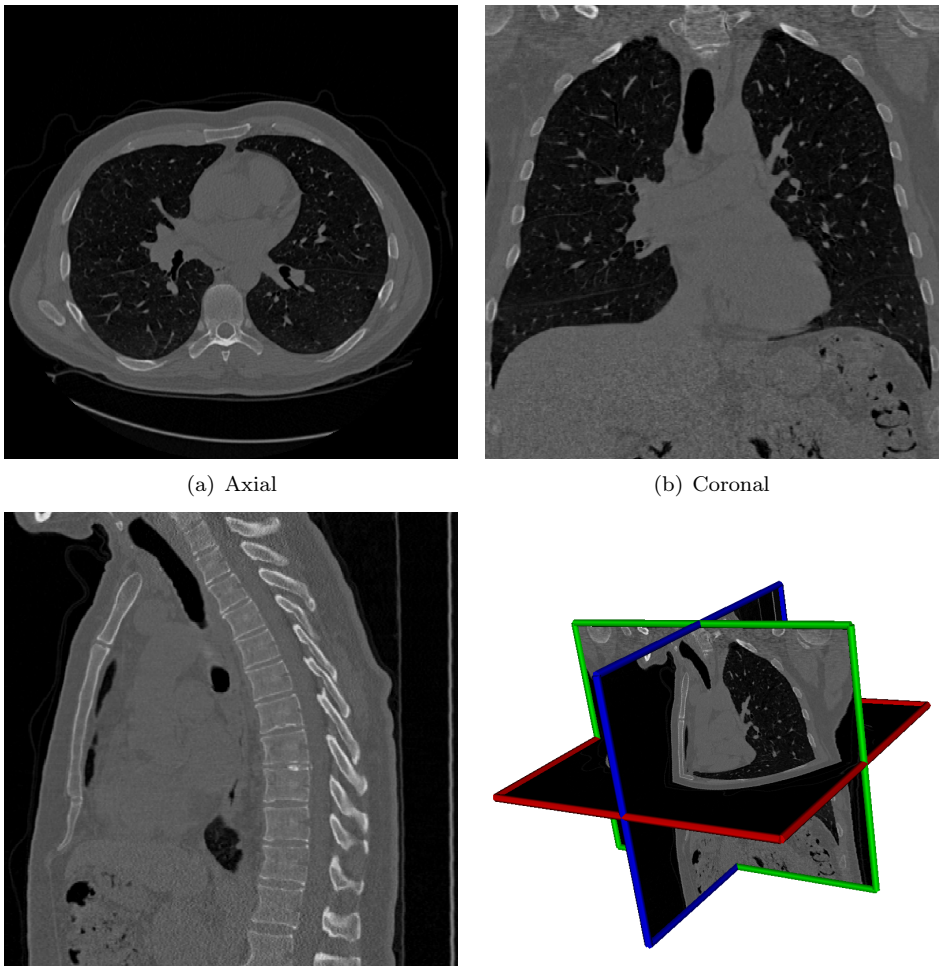
Classification the features are used to assign the sensed objects to a category.

Post-processing other considerations can be taken into account, such as the context.

The methods presented in this thesis are roughly related to this organization in the following way: from a CT image, or CT slices (sensing, the images are viewed as the signal acquired from the subjects by CT scanning), the lung fields are segmented (segmentation) and regions of interest within the lung fields are characterized using texture descriptors (feature extraction) that are fed to a classifier that outputs a probability of disease for the entire CT image (classification). No post-processing is performed, but this could be incorporating other information, e.g., available risk factors such as smoking status.

A central aspect of pattern recognition is the object representation. The classical approach is to measure certain features from the objects, this is essentially the feature extraction step described above, and to represent the objects as feature vectors, or points, in feature vector space where each dimension corresponds to one of the

[1] A pack year is defined as smoking 20 cigarettes a day for one year.



(a) Axial (b) Coronal

(c) Sagittal (d) Three-dimensional rendering

Figure 1.4: An example CT image from the DLCST database. The three views are: (a) axial, (b) coronal, and (c) sagittal view. (d) A three-dimensional rendering of the three body planes.

features measured [24, 43]. Classifiers are then applied in this representation. In this approach, finding good features is important. Alternatively, one can base the classification on direct comparisons between objects. The k'th nearest neighbor (kNN) rule is an example of a classifier that can work directly on object comparisons [24, 43]. However, vector space methods can also be applied to classify objects based on object comparisons. This can be done using a distance representation of the objects, also called the dissimilarity representation approach to pattern recognition [25, 68]. Here, a feature vector space is obtained from the pair-wise object comparisons, or dissimilarities. In this approach, finding a good object dissimilarity measure is important. This thesis takes the second approach to representation. The objects considered are either ROIs or complete CT images, and the dissimilarity between two objects of either type is based on comparing texture. In the case of ROIs, this is simply the textural dissimilarity between the two ROIs being compared. For CT images, textural dissimilarities between ROIs from the two images being compared are combined into an overall CT image dissimilarity.

There exist no generally agreed upon definition of what *texture* is, and attempts at definitions in the literature often depend on the particular application at hand [99]. We will not attempt to provide a definition of texture. However, when referring to texture in this thesis, we loosely mean: *a distribution of measure(s) of local structure, as measured in a local region.* In the extreme case, the measure is the voxel values, and we simply have the histogram of intensity values estimated in a ROI. We would like to highlight one text book definition, however:

> *What is texture? Texture is the variation of data at scales smaller than the scales of interest.* [72]

In the context of analyzing COPD in CT images of lungs by characterizing the lung tissue texture, structures such as the lungs, the lobes, or the pulmonary segments could be seen as the scale of interest. However, since milder stages of emphysema are localized [108], we assume ROIs within the lungs of size smaller than the pulmonary segments to be the scale of interest. The data variation within these ROIs, such as black blobs of varying sizes, is considered as texture.

1.3 Outline of this thesis

The main content of this thesis is presented in five chapters. Chapter 2 presents a trainable texture-based measure for emphysema quantification in CT slices that is trained on manually annotated ROI examples. Chapter 3 extends this to COPD quantification in volumetric CT images and to using PFTs as labels, thereby completely avoiding human input in the training phase. Chapter 4 investigates ROI classification by applying a classifier in a dissimilarity representation of the ROIs, instead of using a kNN classifier to classify ROIs based on textural dissimilarity, as is done in Chapters 2 and 3. A method for classifying sets of objects based on set dissimilarities is presented in Chapter 5, and the same method is applied to classify volumetric CT images in Chapter 6. Here, the classification is based on a CT image dissimilarity

measure that uses the textural dissimilarity between ROIs within the images being compared. The last chapter, Chapter 7, summarizes the thesis and provides a general discussion of the thesis content. Please refer to this chapter for the full summary of the five main chapters, i.e., Chapters 2 to 6.

1.4 Main contributions

The main contributions of this thesis are:

1. Application of local binary patterns (LBPs), a state-of-the-art texture descriptor, to abnormal lung tissue classification, more specifically emphysema classification (Chapter 2).

2. Classification of abnormal lung tissue using full filter response histograms. This is in contrast to using measures computed from the histograms, which is the general trend in texture-based lung tissue classification (Chapters 2, 3, and 6).

3. Texture-based quantitative measures of chronic obstructive pulmonary disease (COPD) using a classifier that is trained on computed tomography (CT) image regions of interest (ROIs) labeled according to the lung function of the subjects. Hereby, manual annotation is completely avoided (Chapters 3 and 6).

4. Conducting the, to our knowledge, largest study of texture-based quantification of COPD in CT images to this date (Chapters 3 and 6).

5. Exploration of dissimilarity approaches for emphysema texture classification, both using a k nearest neighbor (kNN) classifier and in a dissimilarity representation approach (Chapter 4).

6. Dissimilarity-based multiple instance learning (MIL). A novel algorithm for solving the MIL problem is proposed (Chapter 5).

7. Classifying CT images directly based on image dissimilarity (Chapter 6).

Chapter 2

Emphysema Quantification Using Texture

This chapter is based on the manuscript "Quantitative Analysis of Pulmonary Emphysema Using Local Binary Patterns," by L. Sørensen, S. B. Shaker, and M. de Bruijne, published in *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 559–569, 2010.

Abstract We aim at improving quantitative measures of emphysema in computed tomography (CT) images of the lungs. Current standard measures, such as the relative area of emphysema (RA), rely on a single intensity threshold on individual pixels, thus ignoring any interrelations between pixels. Texture analysis allows for a much richer representation that also takes the local structure around pixels into account.

This chapter presents a texture classification based system for emphysema quantification in CT images. Measures of emphysema severity are obtained by fusing pixel posterior probabilities output by a classifier. Local binary patterns (LBP) are used as texture features, and joint LBP and intensity histograms are used for characterizing regions of interest (ROI)s. Classification is then performed using a k nearest neighbor classifier with a histogram dissimilarity measure as distance.

A 95.2% classification accuracy was achieved on a set of 168 manually annotated ROIs comprising the three classes: normal tissue, centrilobular emphysema, and paraseptal emphysema. The measured emphysema severity was in good agreement with a pulmonary function test (PFT) achieving correlation coefficients of up to $|r| = 0.79$ in 39 subjects. The results were compared to RA and to a Gaussian filter bank, and the texture based measures correlated significantly better with PFT than RA did.

2.1 Introduction

Chronic obstructive pulmonary disease (COPD) is a growing health problem worldwide. In the United States alone, it is the fourth leading cause of morbidity and mortality, and it is estimated to become the fifth most burdening disease worldwide by 2020 [77]. COPD is a chronic lung disease characterized by limitation of airflow. It comprises two components: small airway disease and emphysema, which is characterized by gradual loss of lung tissue. Detection and quantification of emphysema is important, since it is thought to be the main cause of shortness of breath and disability in COPD.

The primary diagnostic tool for COPD is spirometry by which various pulmonary function tests (PFT)s are performed [77]. However, PFTs have a low sensitivity to emphysema and are not capable of detecting early stages of COPD [39]. Another diagnostic tool that is gaining more and more attention is computed tomography (CT) imaging. CT is a sensitive method for diagnosing emphysema, assessing its severity, and determining its subtype, and both visual and quantitative CT assessment are closely correlated with the pathological extent of emphysema [58].

In this chapter, we focus on the assessment of emphysema in CT images. Emphysema lesions, or bullae, are visible in CT images as areas of abnormally low attenuation values close to that of air. In CT, emphysema can be classified into three subtypes, or patterns, and we will adopt the naming and definitions used in Webb *et al.* [108]. These subtypes are the following: centrilobular emphysema (CLE), defined as multiple small low-attenuation areas; paraseptal emphysema (PSE), defined as multiple low-attenuation areas in a single layer along the pleura often surrounded by interlobular septa that is visible as thin white walls; and panlobular emphysema (PLE), defined as a low-attenuation lung with fewer and smaller pulmonary vessels. Examples of CLE and PSE, as well as normal tissue (NT), are shown in Fig. 2.1.

Common computerized approaches to emphysema quantification in CT are based on the histogram of CT attenuation values, and different quantitative measures of the degree of emphysema can be derived from this histogram. The most common measure is the relative area of emphysema (RA), also referred to as emphysema index or density mask [58], which measures the relative amount of lung parenchyma pixels that have attenuation values below a certain threshold. Usually, thresholds in the range -856 to -960 Hounsfield units (HU) are used. Measures based on the attenuation histogram disregard the information present in the morphology of the emphysema subtypes such as shape and size distribution of bullae. This was exemplified in a recent clinical study that reported discrepancies between visual scoring and RA for assessing the craniocaudal distribution of the three emphysema subtypes [94].

One way to objectively characterize the emphysema morphology is to describe the local image structure using texture analysis techniques [57, 99]. Uppaluri *et al.* introduced the idea of classifying emphysema in lung CT images using texture features [103]. Several authors followed this idea and classified regions of interest (ROI)s of various lung disease patterns using different texture features, mostly measures on gray-level co-occurrence matrices (GLCM), gray-level run-length matrices (GLRLM),

and on the attenuation histogram, and different classifiers [12, 19, 31, 65, 74, 75, 85, 87, 102, 111]. Other examples of texture features used in the lung tissue classification literature are: the gray-level difference method [74, 75]; discrete wavelet frame decomposition using third order B-splines [19]; convolving with partial derivatives of the Gaussian and the Laplacian of the Gaussian [85, 87]; gradient magnitude [65]; and fractal dimension [102, 103, 111]. In some cases, shape, or geometric, measures are also included in conjunction with the texture features [31, 65, 85]. Most works use a mix of rotation invariant and rotation variant texture features, whereas the texture features used in this chapter are solely rotation invariant.

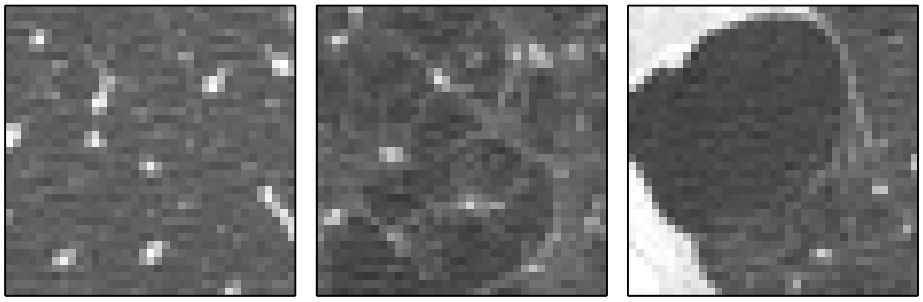
Most of the work on lung texture classification have one or several explicit emphysema classes [12, 19, 31, 65, 74, 75, 102, 103, 111]. Multiple emphysema classes are defined by sub-dividing according to disease severity [65, 111] or emphysema morphology [12, 74, 75]. Chabat *et al.* discriminate between CLE and PLE [12] whereas Prasad *et al.* distinguish between different stages of emphysema, ranging from diffuse to bullous emphysema [74, 75]. The work described in this chapter has two emphysema classes defined based on morphology, namely CLE and PSE. PLE is not considered since only 2 out of 39 subjects had PLE as leading pattern in the data used in the experiments. The data comes from a population of (ex-)smokers, and PLE is known to be more prevalent in subjects with α_1-antitrypsin deficiency than in subjects with smoking-related COPD [39].

A trained classifier can be used for quantification by classifying all pixels in the lung field. In [31, 65, 74, 75, 85, 102, 111] the full lung is classified either by labeling complete ROIs [85, 102, 111] or by labeling individual pixels [31, 65, 74, 75]. Xu *et al.* report the percentage of different disease patterns present in a few subjects, but these quantitative measures are not evaluated further [111]. Park *et al.* quantify emphysema by a weighted sum of relative emphysema class areas [65], and it is to our knowledge the only emphysema based quantitative study on a group of subjects in the lung CT texture analysis literature.

This chapter proposes two new ideas in the area of lung texture analysis in CT images. The specific application is emphysema quantification, but the ideas are also applicable to other lung disease patterns.

The first idea is to use local binary patterns (LBP) originally formulated by Ojala *et al.* [61] as lung texture features. LBP unify structural and statistical information by a histogram of LBP codes that correspond to micro-structures in the image at different scales. LBP have shown promising results in various applications in computer vision and have successfully been applied in a small number of other medical image analysis tasks, e.g., in mammographic mass detection [62] and magnetic resonance image analysis of the brain [101]. In [92], we showed that histogram dissimilarity measures between LBP feature histograms in a k nearest neighbor (kNN) classifier [43] can discriminate between emphysematous and normal tissue.

The second idea is to fuse the posterior probabilities obtained from a classification of all pixels in the lung field into quantitative measures of emphysema severity. Texture based classification allows for quantification of different emphysema subtypes, which may be important in phenotyping emphysema for increased understanding of



(a) (b) (c)

Figure 2.1: Examples of different lung tissue patterns in CT shown with the window setting −600/1500 HU [108]. (a) Normal tissue (NT). (b) Centrilobular emphysema (CLE). (c) Paraseptal emphysema (PSE). The white area in the left part of image (c) is the exterior of the lung.

COPD. Further, texture features may be less influenced by inspiration level and noise compared to, e.g., RA, which uses intensity in single pixels. In [91], we showed that this approach agrees well with the outcome of PFTs, achieving a significant correlation. Two fusion schemes are considered in this chapter; mean class posterior (MCP) and relative class area (RCA). The second fusion scheme, RCA, is related to the fusion scheme in [65] that uses a weighted sum of relative class areas. The difference is that we consider each relative class area individually.

The proposed system is evaluated in two ways; ROI classification and emphysema quantification on subject level. A data set comprising 2D high resolution CT (HRCT) slices with manually annotated ROIs is used for these purposes. The LBP features are compared to two other sets of features, one based on a Gaussian filter bank (GFB) and one comprising measures on GLCM, GLRLM, and the attenuation histogram.

2.2 Methods

The proposed system for emphysema quantification relies on texture classification in local ROIs in the CT images. Three types of texture features are considered, LBP, GFB, and a set of features based on GLCM, GLRLM, and the attenuation histogram. Section 2.2.1 describes LBP, and Section 2.2.2 describes GFB. Measures on GLCM and GLRLM are the most commonly used features in lung texture classification, and they are therefore not described in detail here. We refer to [41, 57] for a detailed description and to [12, 65, 74, 75, 103] for examples of applications. Section 2.2.3 describes how the texture in the ROIs is characterized by computing distributions of features, or feature histograms, and Section 2.2.4 presents a combined measure of histogram dissimilarity between the feature histograms, used to discriminate ROIs with a kNN classifier. Finally, Section 2.2.5 describes how emphysema is quantified in the CT images by fusing pixel posterior probabilities output by a kNN classifier

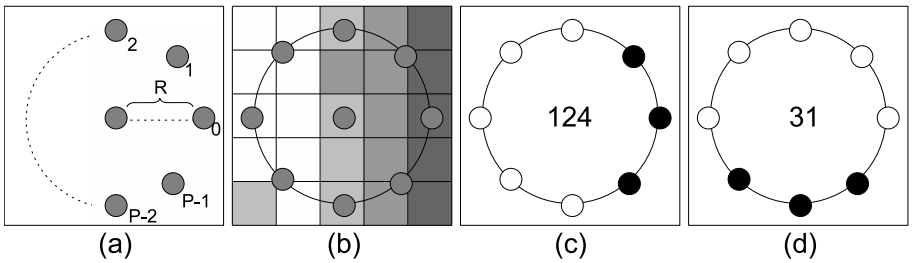


Figure 2.2: Illustration of LBP. (a) The filter is defined by two parameters; the circle radius R and the number of samples P on the circle. (b) Local structure is measured w.r.t. a given pixel by placing the center of the circle in the position of that pixel. (c) The samples on the circle are binarized by thresholding with the intensity in the center pixel as threshold value. Black is zero and white is one. The example image shown in (b) has an LBP code of 124. (d) Rotating the example image in (b) ninety degrees clock-wise reduces the LBP code to 31 which is the smallest possible code for this binary pattern. This principle is used to achieve rotation invariance.

trained on a small set of ROIs.

2.2.1 Local binary patterns

LBP were originally proposed by Ojala *et al.* as a gray-scale invariant measure for characterizing local structure in a 3×3 pixel neighborhood [60]. Later, a more general formulation was proposed that further allowed for multi-resolution analysis and rotation invariance [61]. We use the formulation given in [61]. The LBP are obtained by thresholding samples in a local neighborhood with respect to the center pixel intensity and is given by

$$LBP(\mathbf{x}; R, P) = \sum_{p=0}^{P-1} H(I(\mathbf{x}_p) - I(\mathbf{x}))2^p \tag{2.1}$$

where I is an image, $\mathbf{x}$ is the center pixel, $\mathbf{x}_p = [-R\sin(2\pi p/P), R\cos(2\pi p/P)]^T + \mathbf{x}$ are P local samples taken at a radius R around $\mathbf{x}$, and $H(\cdot)$ is the Heaviside function. As long as the relative ordering among the gray-scale values in the samples does not change, the output of (2.1) stays the same; hence, LBP are invariant to any monotonic gray-scale transformation. The application of the LBP filter is illustrated in Fig. 2.2. Note that, by choosing a fixed sample position on the circle as the "leading bit", in this case the right-most sample, the thresholded samples can be interpreted as bits, and a P bit binary number can be computed.

LBP measure the local structure by assigning unique identifiers, the binary number, to various micro-structures in the image. Thus, LBP capture many structures in one unified framework. In the example in Fig. 2.2(b), the local structure is a vertical edge with a leftward intensity gradient. Other micro-structures are assigned different

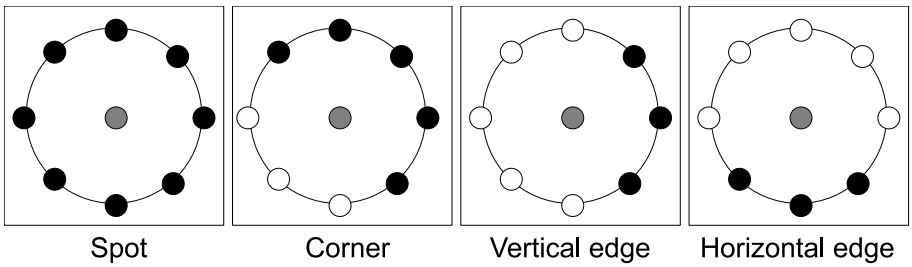


Figure 2.3: Various micro-structures measured by LBP. The gray circle indicates the center pixel. Black and white circles are binarized samples; black is zero and white is one.

LBP codes, e.g., corners and spots as illustrated in Fig. 2.3. By varying the radius R and the number of samples P, the structures are measured at different scales, and LBP allows for measuring large scale structures without smoothing effects, as is, e.g., the case for Gaussian based filters. We expect emphysematous tissue to contain more edges and homogeneous dark areas compared to normal, healthy tissue. Further, the micro-structures are expected to exist at different scales and frequencies according to the severity of the disease state.

Rotation invariant LBP are achieved by "rotating the circle" until the lowest possible binary number is found

$$LBP^{ri}(\mathbf{x}; R, P) = \min_i(ROR(LBP(\mathbf{x}; R, P), i)) \tag{2.2}$$

for $i = 0, \ldots, P-1$. $ROR(b, i)$ performs i circular bit-wise right shifts on the P-bit binary number b. When using (2.2), the horizontal edge and the vertical edge in Fig. 2.3 are assigned the same LBP code, namely 31. We will use the LBP formulation in (2.2) in all experiments.

2.2.2 Gaussian filter bank

The second type of texture features are computed using a rotation invariant GFB and are based on convolving the image with the Gaussian function

$$G(\mathbf{x}; \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(- \frac{||\mathbf{x}||_2^2}{2\sigma^2} \right) \tag{2.3}$$

where σ is the standard deviation, or scale.

Sluimer *et al.* used a similar GFB comprising both rotation variant and invariant filters [85, 87]. Since rotation invariant LBP are used in this chapter, the GFB we compare with consists of the rotation invariant filters from [85, 87]; Gaussian and Laplacian of the Gaussian, augmented with two more rotation invariant filters; gradient magnitude, which is also used on the original data in [65], and Gaussian curvature.

Letting L_x and L_y denote the first order derivatives of the convolved image $L = I * G(\mathbf{x}; \sigma)$, and L_{xx}, L_{yy} and L_{xy} denote the second order derivatives, the four base filters in the GFB are as follows: the Gaussian function (2.3) itself, the Laplacian of the Gaussian

$$\nabla^2 G(\mathbf{x}; \sigma) = L_{xx} + L_{yy}, \tag{2.4}$$

gradient magnitude

$$||\nabla G(\mathbf{x}; \sigma)||_2 = \sqrt{L_x^2 + L_y^2}, \tag{2.5}$$

and Gaussian curvature

$$K(\mathbf{x}; \sigma) = L_{xx} L_{yy} - L_{xy}^2. \tag{2.6}$$

2.2.3 Feature histograms

Based on the feature values in an ROI, obtained either by computing rotation invariant LBP (2.2) in all pixels in the ROI or by applying one of the GFB filters (2.3), (2.4), (2.5), or (2.6), a feature histogram, $f(\text{ROI})$, is computed.

For LBP, the computed LBP codes are directly accumulated into a histogram with the number of bins determined by the number of samples P. In the case of GFB, we employ an adaptive binning principle similar to that of [60]; the total feature distribution across all ROIs in the training set is made approximately uniform. Consequently, densely populated areas in feature space are quantized with a high resolution while sparse areas are quantized with a low resolution. The number of bins is set to $\lfloor \sqrt[3]{N_p} \rfloor$, where N_p is the number of pixels in the ROI.

As noted previously, LBP are invariant to any monotonic gray-scale transformation of the image. This is, however, not always desirable when dealing with CT images, where values are measurements of a physical property of the tissue displayed [87]. Therefore, we include intensity information in the feature histogram by forming the joint histogram between the LBP codes and the intensities in the center pixels. The intensities are binned using the same adaptive principle as used for the GFB filter values [60].

Examples of feature histograms computed from the three different ROIs in Fig. 2.1 are shown in Fig. 2.4. Only few bins contain the mass in the LBP histograms, and these bins correspond to different micro-structures such as edges, corners, and spots, as indicated with arrows in Fig. 2.4(d). One of the discriminating bins when comparing the NT ROI to the CLE ROI is the edge bin as expected, see Fig. 2.4(d) and Fig. 2.4(e). The joint LBP and intensity histogram captures information about at which intensities the different micro-structures reside, thus improving discrimination of the NT ROI from the CLE ROI when, e.g., looking at the edge bin in Fig. 2.4(a) and Fig. 2.4(b).

2.2.4 Classifier

The feature histograms are used to classify ROIs or center pixels of ROIs. For this purpose, we use the kNN classifier [43] with the distance between two ROIs being

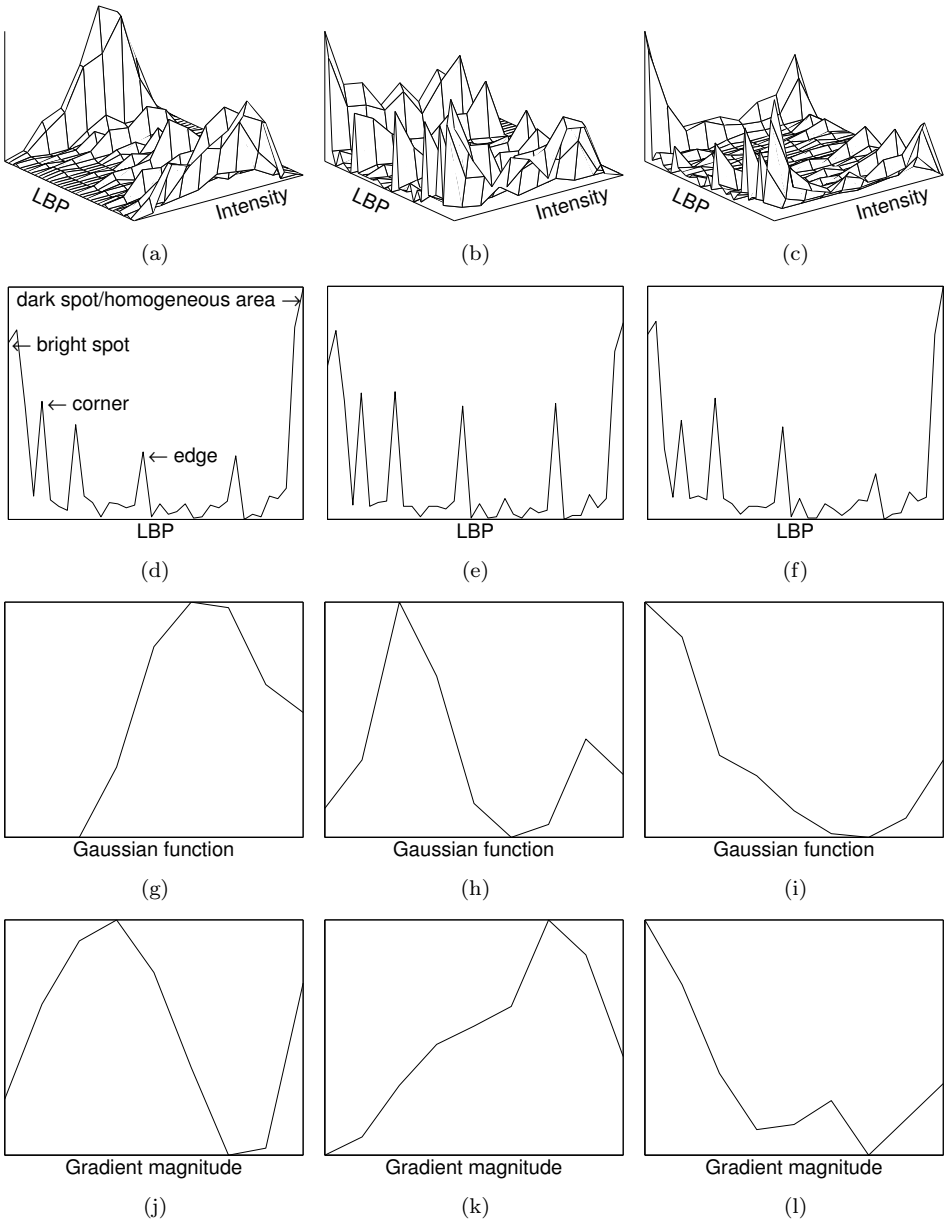


Figure 2.4: Examples of feature histograms. (a, d, g, j) Are computed from the NT ROI in Fig. 2.1(a), (b, e, h, k) from the CLE ROI in Fig. 2.1(b), and (c, f, i, l) from the PSE ROI in Fig. 2.1(c). (a-c) Joint LBP and intensity histograms for $R = 1$ and $P = 8$. (d-f) LBP histograms for $R = 1$ and $P = 8$. (g-i) Gaussian function filter response histograms for $\sigma = 0.5$. (j-l) Gradient magnitude filter response histograms for $\sigma = 1$.

a combined histogram dissimilarity between feature histograms. kNN is the natural classifier of choice when working in a distance representation of objects, and it is also the classifier employed in the LBP literature [60, 61]. Further, kNN is a non-parametric classifier and therefore able to handle multi-modal class distributions in feature space, which might be the case in lung texture classification. For example, an emphysema class containing samples from different disease stages might contain patterns of varying bullae sizes and varying number of edges, giving rise to a multi-modal class distribution.

Each ROI is represented by a set of feature histograms. When using LBP, the set comprises feature histograms that are measured with different radii for multi-scale analysis. When using GFB, the set comprises feature histograms measured with different filters at different scales. Dissimilarities between ROIs are expressed as dissimilarities between feature histogram sets

$$D_{set}(\mathrm{ROI}_j, \mathrm{ROI}_k) = \sum_{i}^{N_f} D(f_i(\mathrm{ROI}_j), f_i(\mathrm{ROI}_k))$$

where N_f is the number of histograms, $D(\cdot,\cdot)$ is a histogram dissimilarity measure, and $f_i(\cdot)$ are individual feature histograms. In this chapter, we use negated histogram intersection [95] as histogram dissimilarity measure

$$D(H, K) = -\sum_{i=1}^{N_b} \min(H_i, K_i)$$

where H and K are histograms each with N_b bins. In this chapter, all feature histograms are normalized to sum to one, thus $D(\cdot,\cdot) \in [-1, 0]$.

We use a posterior probability estimator for the kNN classifier that includes distances to prototypes in the estimation. A principle similar to [27] is employed; the estimation is based on the distance to the n'th nearest prototype of each class where n is the number of prototypes of the majority class within the k nearest neighbors of $\mathbf{x}$. The posterior probability of class ω_i given pixel $\mathbf{x}$ is therefore given by

$$P(\omega_i|\mathbf{x}) = \frac{|D_{set}(\mathrm{ROI}(\mathbf{x}), \mathrm{ROI}_n^{\omega_i})|}{\sum_{j=1}^{N_c} |D_{set}(\mathrm{ROI}(\mathbf{x}), \mathrm{ROI}_n^{\omega_j})|} \tag{2.7}$$

where N_c is the number of classes, $\mathrm{ROI}_n^{\omega_i}$ is the n'th nearest prototype of class ω_i, and $\mathrm{ROI}(\mathbf{x})$ is the ROI centered on $\mathbf{x}$.

2.2.5 Emphysema quantification

Prior to classification of the lung field, the lung parenchyma pixels are segmented in the HRCT slice using a combination of thresholding and connected component analysis. Manual editing was needed afterwards in one third of the cases and required simple outlining of a few of the larger airways. In principle, automated methods such as [42, 54] could be used here instead. We denote the obtained segmentation S. Each

segmented lung parenchyma pixel is classified by classifying the ROI centered on the pixel.

It should be noted that pixels that are not part of the lung segmentation S are not classified, but they can still contribute to the classification. For example, part of the exterior of the lung is in the local neighborhood when classifying a pixel at the border of the lung. In this way, all potentially relevant structural information is incorporated, such as proximity to the border of the lung or to the large vessels and airways.

The pixel probabilities are fused to obtain one measure for the complete lung field that can be used for emphysema quantification. There are several ways of doing this, e.g., averaging, voting, or the maximum rule [55]. In this chapter, we evaluate averaging of soft and hard classification results. The considered quantitative measures for emphysema are the mean class posterior (MCP_{ω_i}) and the relative class area (RCA_{ω_i}). MCP_{ω_i} is given by

$$\mathrm{MCP}_{\omega_i} = \frac{1}{|S|} \sum_{\mathbf{x}_j \in S} P(\omega_i|\mathbf{x}_j) \tag{2.8}$$

where $|S|$ is the number of lung parenchyma pixels in segmentation S and $P(\omega_i|\mathbf{x}_j)$ is obtained using (2.7). RCA_{ω_i} is given by

$$\mathrm{RCA}_{\omega_i} = \frac{1}{|S|} \sum_{\mathbf{x}_j \in S} \delta(\arg\max_c P(\omega_c|\mathbf{x}_j) - i) \tag{2.9}$$

where $\delta(\cdot)$ denotes the Kronecker delta function.

2.3 Experiments and results

2.3.1 Data

The data comes from an exploratory study carried out at the Department of Respiratory Medicine, Gentofte University Hospital [84] and consist of CT images of the thorax acquired using General Electric equipment (LightSpeed QX/i; GE Medical Systems, Milwaukee, WI, USA) with four detector rows. A total of 117 HRCT slices were acquired by scanning 39 subjects in the upper, middle, and lower lung. The CT scanning was performed using the following parameters: in-plane resolution 0.78×0.78 mm, 1.25 mm slice thickness, tube voltage 140 kV, and tube current 200 mAs. The slices were reconstructed using a high spatial resolution (bone) algorithm.

Prior to CT imaging, the subjects underwent PFTs, and both the forced vital capacity (FVC) and the forced expiratory volume in one second (FEV_1) were measured [76]. FEV_1 is adjusted for age, sex, and height by dividing with a predicted value according to these three parameters, thereby obtaining FEV_1%pred.

The 39 subjects were divided into three groups: 9 healthy lifelong non-smokers (referred to as never-smokers), 10 smokers without COPD (referred to as healthy

Table 2.1: Group characteristics reported as mean values, with standard deviation in parentheses and range in square brackets. n is the number of subjects in a group.

Group	**Age**	**FEV$_1$**	**FEV$_1$%pred**	**FEV$_1$/FVC**
Never-	59	3.15	103	80
smokers	(9)	(0.77)	(9)	(4)
$n = 9$	[47-73]	[2.02-4.08]	[93-121]	[76-89]
Healthy	58	2.90	101	78
smokers	(10)	(0.47)	(8)	(5)
$n = 10$	[47-73]	[1.95-3.60]	[85-113]	[68-78]
COPD	64	1.62	57	54
smokers	(8)	(0.57)	(12)	(7)
$n = 20$	[49-80]	[0.94-2.73]	[37-76]	[42-67]

smokers), and 20 smokers diagnosed with moderate or severe COPD (referred to as COPD smokers). The COPD diagnosis was based on the recorded PFTs and done according to the Global Initiative for Chronic Obstructive Lung Disease criteria [77] as follows: no COPD, defined as $\text{FEV}_1/\text{FVC} \geq 0.7$ and $\text{FEV}_1\%\text{pred} \geq 80\%$; moderate to severe COPD, defined as $\text{FEV}_1/\text{FVC} < 0.7$ and $30\% \leq \text{FEV}_1\%\text{pred} < 80\%$. Of the 39 subjects, 19 were women and 20 were men. Table 2.1 summarizes the characteristics of the three groups.

An experienced chest radiologist and a CT experienced pulmonologist each assessed the leading pattern, either NT, CLE, PSE, or PLE, in each of the 117 slices. Overall, the observers agreed in 53% of the slices, and they agreed on the emphysema class in 60% of slices where both decided on an emphysema pattern.

168 non-overlapping ROIs were annotated manually in 25 of the subjects, representing the three classes: NT (59 observations), CLE (50 observations), and PSE (59 observations). The NT ROIs were annotated in never-smokers, and the CLE and PSE ROIs were annotated in healthy smokers and COPD smokers within the area(s) of the leading emphysema pattern by approximately marking the center pixel of the emphysematous area. Square ROIs of a given width centered on the marked pixel were subsequently extracted. PLE was excluded due to under-representation in the data, only two subjects had PLE as leading pattern. Therefore, we are dealing with the three classes $\omega_i \in \{\text{NT}, \text{CLE}, \text{PSE}\}$ in all the experiments.

2.3.2 Feature and parameter selection

When using the GFB, feature selection is applied using the sequential forward selection algorithm [43] for deciding which filters at which scales to include. When using LBP, several combinations of radii for multi-resolution analysis are evaluated. In both approaches, different k's in the kNN classifier as well as different ROI sizes

are evaluated during training. In all cases, parameters and feature sets are optimized based on validation classification accuracy.

2.3.3 Classification of ROIs

Classification performance is evaluated by leave-one-subejct-out error estimation on the set of manually annotated ROIs. Six different approaches are evaluated and compared.

The ROIs are represented as points in a feature space, with all features standardized to unit variance, in the first two approaches, and Euclidean distance in the feature space is used in the kNN classifier:

1) GFB1 The feature vector consists of the first four central moments computed from histograms of GFB filter responses. Standard histograms are used instead of applying the adaptive binning approach described in Section 2.2.3, and the four filters described in Section 2.2.2 are used, resulting in a $16 \times$ scales dimensional feature vector. This set of features resembles the features used in [85, 87].

2) Intensity, Co-occurrence, and Run-length (ICR) The feature vector consists of the following features: the first four central moments of the intensity histogram; the gray-level co-occurrence matrix (GLCM) based measures contrast, correlation, energy, entropy, and homogeneity [41, 57]; and the gray-level run-length matrix (GLRLM) based measures short run emphasis, long run emphasis, gray-level nonuniformity, run-length nonuniformity, and run percentage [41, 57]. The resulting feature vector is 14 dimensional. This set of features resembles the features used in [12, 65, 74, 75, 102, 103, 111].

The remaining four approaches all use the methods described in Sections 2.2.3 and 2.2.4 with different feature histograms:

3) INT Intensity histograms.

4) GFB2 GFB filter response histograms.

5) LBP1 Basic rotation invariant LBP histograms.

6) LBP2 Joint 2D LBP and intensity histograms.

In each leave-out trial, all ROIs from one subject are held out and used for testing. The remaining subjects are separated into a training set and a validation set. In this separation, balanced class distributions are ensured by placing half the subjects representing one class in the training set and the rest in the validation set. The optimal parameter setting is learned using the training and validation sets and can differ for each test subject. Subsequently, the ROIs in the test set are classified using the optimal parameter setting and all the ROIs in the training set and validation set as prototypes in the kNN classifier.

In GFB1 and GFB2, the following scales are used for all filters: $\sigma = \{0.5, 1, 2, 4, 8\}$ pixels. In ICR, GLCM and GLRLM are computed using the orientations $\{0, 45, 90, 135\}^\circ$

Table 2.2: ROI classification accuracy and p-value for difference with LBP2 according to McNemar's test

Approach	**Accuracy**	***p*-value**
GFB1	61.3	$< 10^{-4}$
ICR	89.3	0.016
INT	87.5	0.004
GFB2	94.0	0.724
LBP1	79.2	$< 10^{-4}$
LBP2	95.2	-

and the lengths $\{1, 2, \ldots, 5\}$ pixels, and the following binnings of intensity values are evaluated: $\{16, 32, 64\}$ number of bins. The GLCMs are symmetric and mean GLCM measures across orientation and length are used [12, 65]. GLRLM are computed using the Gray Level Run Length Matrix Toolbox [109], and mean GLRLM measures across orientation are used. In LBP1 and LBP2, the following radii and corresponding number of samples are used: $R = \{1, 2\}$ pixels and $P = \{8, 16\}$ samples. Common parameters considered for all six approaches are as follows: ROI size $= \{31 \times 31, 41 \times 41, 51 \times 51\}$ pixels and number of neighbors in the kNN classifier $k = \{1, 2, \ldots, 10\}$.

The estimated classification accuracies of the six approaches are summarized in Table 2.2. LBP2 performs best, achieving a classification accuracy of 95.2%. However, it is not significantly different ($p = 0.72$), according to a McNemar's test [20], from the second best approach, GFB2, which achieves an accuracy of 94.0%. As expected, including intensity is important. This is seen in the performance gain between LBP1 and LBP2. In fact, intensity alone performs better than LBP alone, as seen when comparing INT to LBP1. LBP2 performs significantly better than the four approaches GFB1, ICR, INT, and LBP1 ($p < 0.05$). We will focus on the two best performing approaches, GFB2 and LBP2, in the remaining part of Section 2.3.

The confusion matrices in Table 2.3 show that LBP2 and GFB2 generally agree on the class labels. Further, GFB2 never mistakes the emphysema classes, and LBP2 only labels a PSE pattern as CLE once. The agreement between the two approaches is further investigated in Section 2.3.4.

The parameter settings and filters that were most often selected in the leave-one-subject-out error estimation, for LBP2 and GFB2, are shown in Table 2.4 and Table 2.5, respectively. The tendency is small scale features, small ROIs, and small k.

2.3.4 Parenchyma classification

In this section, results of applying the trained classifiers to all pixels within the lung fields are compared for LBP2 and GFB2.

Only one parameter setting is considered for each representation based on the most

Table 2.3: Confusion matrices showing the true label (rows) vs. label assigned by the kNN classifier (columns) for the two best performing approaches.

LBP2

	NT	CLE	PSE
NT	55	0	4
CLE	1	49	0
PSE	2	1	56

GFB2

	NT	CLE	PSE
NT	55	0	4
CLE	2	48	0
PSE	4	0	55

Table 2.4: Most frequently selected parameters and filter combinations for LBP2 in the leave-one-subject-out experiments in Section 2.3.3. Only parameters and filters selected in at least 20% of the leave-out trials are shown.

$LBP^{ri}(\mathbf{x}; R = 1, P = 8)$	$LBP^{ri}(\mathbf{x}; R = 2, P = 16)$	$\{LBP^{ri}(\mathbf{x}; R = 1, P = 8), LBP^{ri}(\mathbf{x}; R = 2, P = 16)\}$	$k = 1$	ROI size = 31
56%	20%	24%	96%	96%

Table 2.5: Most frequently selected parameters and filter combinations for GFB2 in the leave-one-subject-out experiments in Section 2.3.3. Only parameters and filters selected in at least 20% of the leave-out trials are shown. Other GFB2 filters that are selected together with the reported GFB2 filter combinations in less than 20% of the individual experiments are not shown.

$\{G(\mathbf{x}; \sigma = 0.5), \|\|\nabla G(\mathbf{x}; \sigma = 1)\|\|_2\}$	$\{G(\mathbf{x}; \sigma = 0.5), \|\|\nabla G(\mathbf{x}; \sigma = 2)\|\|_2\}$	$\{G(\mathbf{x}; \sigma = 0.5), \|\|\nabla G(\mathbf{x}; \sigma = 1)\|\|_2, \|\|\nabla G(\mathbf{x}; \sigma = 2)\|\|_2\}$	$\{G(\mathbf{x}; \sigma = 0.5), \|\|\nabla G(\mathbf{x}; \sigma = 1)\|\|_2, \|\|\nabla G(\mathbf{x}; \sigma = 4)\|\|_2\}$	$k = 1$	$k = 3$	ROI size = 31
92%	28%	24%	28%	64%	20%	96%

Table 2.6: Confusion matrix between GFB2 and LBP2 across all subjects for all lung parenchyma pixels. The numbers reported are in percentage of total number of lung parenchyma pixels.

	LBP2 NT	**LBP2 CLE**	**LBP2 PSE**
GFB2 NT	48.2	1.0	4.1
GFB2 CLE	2.9	16.8	2.5
GFB2 PSE	2.9	0.4	21.3

frequent parameters in Table 2.4 and Table 2.5. For LBP2, we use $\{LBP^{ri}(\mathbf{x}; R = 1, P = 8), k = 1, \text{ROI size} = 31\times31\}$, and for GFB2, we use $\{\{G(\mathbf{x}; \sigma = 0.5), ||\nabla G(\mathbf{x}; \sigma = 1)||_2\}, k = 1, \text{ROI size} = 31 \times 31\}$. The set of annotated ROIs serve as prototypes in the kNN classifier. When classifying the HRCT slices from a particular subject, all the ROI prototypes coming from that same subject are left out in the kNN classifier.

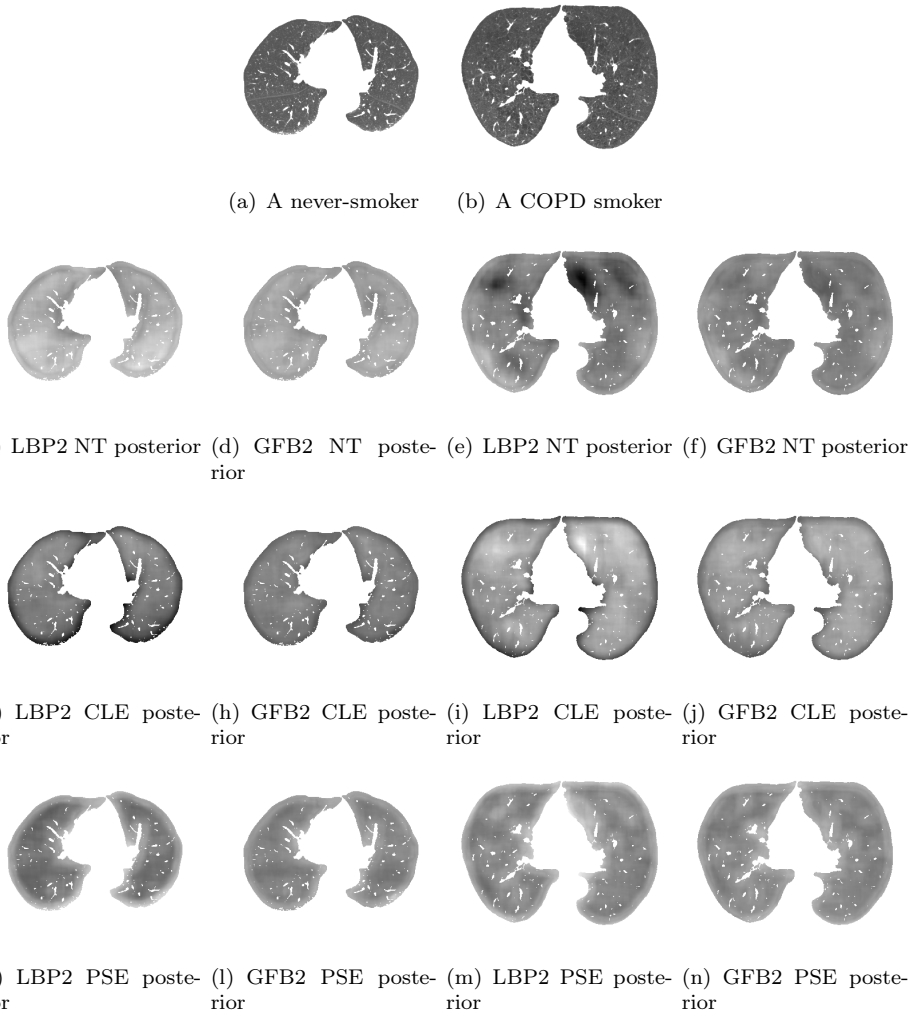
Fig. 2.5 shows examples of the resulting posterior class probabilities assigned by the classifiers in a never-smoker HRCT slice and a COPD smoker HRCT slice. The never-smoker has many high NT probability pixels assigned by both LBP2 and GFB2 as seen in Fig. 2.5(c) and Fig. 2.5(d), whereas the COPD smoker has many high CLE probability pixels and some high PSE probability pixels, see Figs. 2.5(i), 2.5(j), 2.5(m), and 2.5(n). For the shown COPD smoker, the consensus reading of the leading pattern is CLE in all three slices. The LBP2 posterior seems more localized than the GFB2 posterior. See, e.g., the low NT posterior area in the anterior part of the left lung in the slice in Figs. 2.5(e) and 2.5(f) and the high CLE posterior area in the same positions in Figs. 2.5(i) and 2.5(j).

Correlating the class posteriors shows a high degree of agreement between LBP2 and GFB2; $r = 0.93$ ($p < 10^{-4}$) when correlating $P(\text{NT}|\mathbf{x})$ of the two classifiers, $r = 0.94$ ($p < 10^{-4}$) in the case of $P(\text{CLE}|\mathbf{x})$, and $r = 0.91$ ($p < 10^{-4}$) in the case of $P(\text{PSE}|\mathbf{x})$. Further, class label agreements between LBP2 and GFB2 in each lung parenchyma pixel are shown in the confusion matrix in Table 2.6. This result is based on a hard classification obtained by applying the maximum a posteriori rule in each pixel. The two classifiers generally are in good agreement; in 86.3% of the pixels, the two classifiers agree on the class label.

2.3.5 Emphysema quantification

In this section, we evaluate the value of fusing pixel posterior probabilities, computed using the proposed classification system, into a single measure for emphysema.

The full lung classification results of Section 2.3.4 are turned into quantitative measures of emphysema using MCP_{ω_i} according to (2.8) and using RCA_{ω_i} according to (2.9). These measures are computed across the three HRCT slices representing a subject. We evaluate the obtained measures by correlating with FEV_1%pred, which is one of the standard PFTs for diagnosing subjects with COPD [77]. The common CT based measure RA is also included in the evaluation, in this case using a threshold



(a) A never-smoker (b) A COPD smoker

(c) LBP2 NT posterior (d) GFB2 NT posterior (e) LBP2 NT posterior (f) GFB2 NT posterior

(g) LBP2 CLE posterior (h) GFB2 CLE posterior (i) LBP2 CLE posterior (j) GFB2 CLE posterior

(k) LBP2 PSE posterior (l) GFB2 PSE posterior (m) LBP2 PSE posterior (n) GFB2 PSE posterior

Figure 2.5: An HRCT slice from a never-smoker and from a COPD smoker together with posterior probabilities computed in each lung parenchyma pixel. White is high probability and black is low probability. (a, b) Original HRCT slices shown with the window setting $-600/1500$ HU [108]. (c, g, k) LBP2 based posteriors for the never-smoker. (d, h, l) GFB2 based posteriors for the never-smoker. (e, i, m) LBP2 based posteriors for the COPD smoker. (f, j, n) GFB2 based posteriors for the COPD smoker.

Table 2.7: Correlation of CT based emphysema measures with FEV_1%pred. The p-values of the correlations are shown in parentheses.

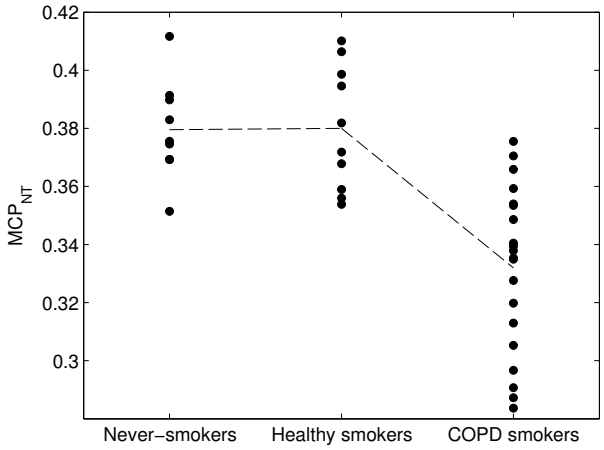
Measure		r
LBP2	MCP_{NT}	0.77 ($p < 10^{-4}$)
	MCP_{CLE}	-0.74 ($p < 10^{-4}$)
	MCP_{PSE}	-0.40 ($p = 0.011$)
	RCA_{NT}	0.77 ($p < 10^{-4}$)
	RCA_{CLE}	-0.78 ($p < 10^{-4}$)
	RCA_{PSE}	-0.66 ($p < 10^{-4}$)
GFB2	MCP_{NT}	0.76 ($p < 10^{-4}$)
	MCP_{CLE}	-0.79 ($p < 10^{-4}$)
	MCP_{PSE}	-0.28 ($p = 0.088$)
	RCA_{NT}	0.75 ($p < 10^{-4}$)
	RCA_{CLE}	-0.74 ($p < 10^{-4}$)
	RCA_{PSE}	-0.66 ($p < 10^{-4}$)
RA_{910}		-0.62 ($p < 10^{-4}$)

of -910 HU [58, 84, 94] (RA_{910}). The results are shown in Table 2.7 where the NT based measures achieve correlation coefficients ranging from $r = 0.75$ to $r = 0.77$. For comparison, RA_{910} correlates significantly worse with FEV_1%pred than the NT based measures do ($p < 0.05$) according to a Hotelling/Williams test [106]. In the Hotelling/Williams test, we correct for the difference in signs.

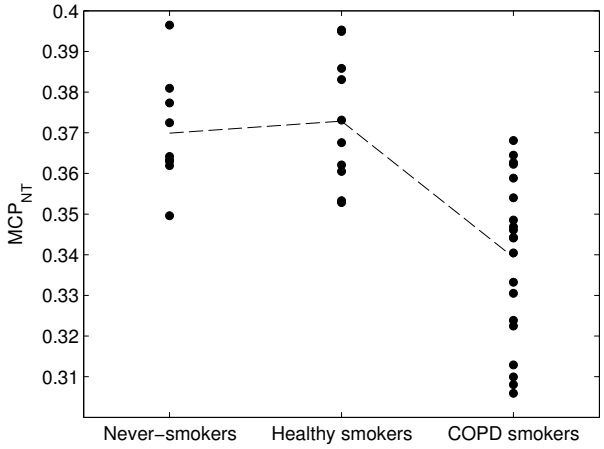
All measures, except LBP2 and GFB2 based MCP_{PSE}, separate the group of COPD smokers from the combined group of never-smokers and healthy smokers according to a rank sum test ($p < 0.05$). The separation can also be seen for LBP2 based MCP_{NT} in Fig. 2.6(a). The figure also shows that the individual features of the joint LBP and intensity feature histogram measure different properties of the parenchyma at subject level. Using intensity alone, i.e., parenchyma density, results in the picture shown in Fig. 2.6(b), and using LBP alone, i.e., parenchyma micro-structures, results in the picture shown in Fig. 2.6(c).

2.4 Discussion and conclusion

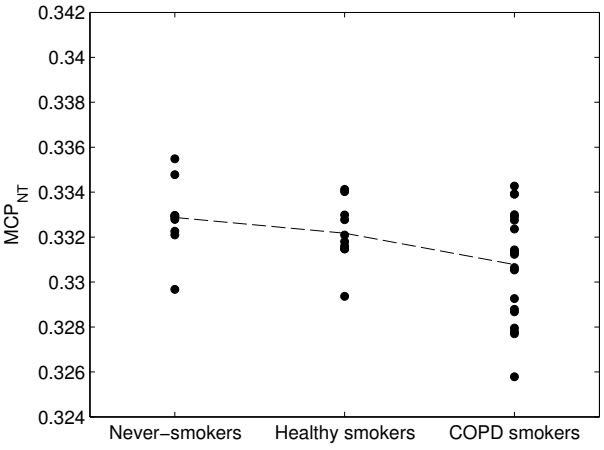
The proposed classification system using LBP2 achieves an ROI classification accuracy of 95.2%, see Table 2.2, with an NT sensitivity and specificity of 97.3% and 93.2% respectively. This is better than using GFB or ICR, and is within the $75 - 100$% range of NT sensitivities and specificities reported in the literature [12, 19, 65, 85, 87, 102, 111]. The experiments revealed that using LBP in isolation does not work



(a) LBP and intensity jointly, i.e., LBP2 based.



(b) Intensity alone, i.e., INT based.



(c) LBP alone, i.e., LBP1 based.

Figure 2.6: MCP_{NT} for all 39 subjects divided on the three groups; never-smokers, healthy smokers, and COPD smokers. The dashed lines are connecting the means of the three groups.

well in the presented application. This was to be expected, since LBP by design are invariant to monotonic intensity transformations and therefore discard the density information contained in the CT image intensities. Including intensity information via the joint LBP and intensity histogram combines complementary information in the form of micro-structures and densities. Hereby it is measured at which densities the different micro-structures reside which improves discrimination considerably. This is illustrated when comparing Figs. 2.4(a) and 2.4(b) to Figs. 2.4(d) and 2.4(e) where the differences between the joint histograms 2.4(a) and 2.4(b) are much more obvious than between the LBP histograms 2.4(d) and 2.4(e). Other feature sets, like GFB, also include intensity and hereby also mix structure and density information. However, obtaining a similar representation to LBP2 in GFB, i.e., joint histograms of structure and density, would require histograms of much higher dimensionality with one dimension for each type of micro-structure such as edge, blob, etc., potentially leading to problems with overfitting in cases of limited number of training samples.

CT based computerized quantitative measures of emphysema are often evaluated by correlating the obtained CT measures with other markers for disease, such as PFTs or plasma biomarkers, in the clinical literature. A few examples of such studies are: [37, 45, 81, 84, 88]. In this chapter, we have performed a similar evaluation and correlated the proposed quantitative measures with another marker for emphysema in Section 2.3.5, namely FEV_1%pred, and in general the correlations were strong, up to $|r| = 0.79$. It is known that PFTs are noisy measurements [22], and that they are affected by other phenomena than emphysema, e.g., inflammation in the airways. Still, some degree of agreement between PFTs and CT based emphysema measurements is expected.

Park *et al.* previously performed emphysema quantification based on a hard classification of the lung parenchyma pixels. A weighted sum of the relative areas of mild end severe emphysema was used, and they reported a correlation of -0.47 with FEV_1%pred [65]. Based on the results of our experiments, nothing conclusive can be stated about fusion of soft versus fusion of hard classifications, i.e., (2.8) versus (2.9). Both methods work well, and as seen in Table 2.7, all the NT based measurements correlate significantly with FEV_1%pred with correlation coefficients in the range $r = 0.75$ to $r = 0.77$. It should be noted that in [65], the correlation between RA, using a threshold of -950 HU, and FEV_1%pred is -0.42. In our data, agreement between CT and PFT generally seems to be better with the correlation between RA_{910} and FEV_1%pred being -0.62. This fact, as well as the fact that we are dealing with a broader range of subjects compared to [65], could explain the difference in correlation level.

Different features may capture different information in the CT images, and though the per pixel posterior probabilities of LBP2 and GFB2 are highly correlated, there may still be something to gain by combining the output of the two classifiers. This was tested by combining the pixel posteriors $P(\omega_i|\mathbf{x})$ of LBP2 and GFB2 with the sum rule and the maximum rule respectively [47], followed by posterior fusion of the combined pixel posteriors. These results did not show any significant improvement in correlation with FEV_1%pred, which indicates that the two classifiers indeed capture

similar information in the CT images.

In this chapter, we have done no preprocessing of the HRCT slices prior to computing feature histograms. Instead, we have relied on the filters to perform the necessary processing, e.g., noise smoothing in the GFB by selecting the appropriate σ or picking up certain micro-structures from the noisy background in LBP by selecting the appropriate radius. It is up to the training procedure to pick these settings.

Further, all available information has been taken into account in the feature histogram estimation by including all pixels in the ROIs instead of first excluding, e.g., the vessels [12, 111] or the airways [111]. Thus, also pixels outside the lung fields and pixels from non-parenchyma structures within the lungs contribute. This can be seen as an implicit way of encoding context information in the feature histograms, with the position being "near the border of the lung" or "near the hilar area". It may lead to slight overestimation of PSE at the border, see Fig. 2.5, but in practice, this should not be a real problem as long as the prototype set contains samples at the border representing all classes. In our data, NT and CLE ROIs were mainly annotated in the central parts of the lungs. Nevertheless, the proposed classification system is capable of discriminating between normal tissue and emphysematous tissue within the lung, as seen in the confusion matrices for LBP2 and GFB2 in Table 2.3.

Fig. 2.6(a) reveals that very similar measures are obtained for the never-smokers and the healthy smokers. One might expect the healthy smokers measures to be slightly lower than the never-smokers on the $\mathrm{MCP}_{\mathrm{NT}}$ scale due to early stages of emphysema not yet detectable by PFTs. Basing the measurements only on intensity feature histograms results in the healthy smokers having an even larger probability of NT as seen in Fig. 2.6(b), indicating a difference in density for the two groups. This corresponds well with recent results indicating that lung parenchyma is more dense in healthy smokers than in never-smokers possibly due to smoke induced inflammation [4,88]. On the other hand, basing the measurements solely on LBP feature histograms results in a slight drift downwards as seen in Fig. 2.6(c), suggesting that there may be structural differences that can be captured at an early stage by LBP. As described in Section 2.2.1, LBP are gray-scale invariant and therefore not affected by parenchymal density changes. This also implies that the proposed classification system should be less sensitive to inspiration level as compared to, e.g., RA_{910}.

Basing the discrimination of ROIs on dissimilarities between sets of feature histograms, using a combined histogram dissimilarity directly as distance in a kNN classifier, works well in this setting. Both LBP2 and GFB2 achieve good ROI classification accuracies and high correlations with FEV_1%pred. Using full feature histograms differs from the common approach of using measures derived from feature histograms, such as moments of filter response histograms or GLCM measures, as features in a feature space [12, 19, 31, 65, 74, 75, 85, 87, 102, 103, 111]. Looking at Fig. 2.4, taking only the first four moments of the GFB histograms could potentially discard valuable information about the shape of the histograms such as the presence of multiple mods. A previous comparative study of texture features for classification reported similar findings on two standard texture data sets [60]. In this chapter, we wanted to exploit the full feature histograms and therefore used the kNN classification framework

with histogram dissimilarity as distance, and LBP and GFB were shown to work very well in this setting. It remains of course a possibility that in a different classification scheme, relying on features rather than on dissimilarity measures, a different feature set would perform as good as or even better than LBP. Alternatively, histogram dissimilarities could be applied within the dissimilarity-based classification schemes proposed by Pekalska *et al.* [68].

The experiments carried out in this chapter are all done on HRCT slices, but the general framework could easily be extended to 3D. However, no true extension of rotation invariant LBP to 3D exists. Two approximative extensions of LBP to 3D are presented in [112], with the specific application being temporal texture data. The first approach forms a helical path in the temporal direction. This idea could be applied in volumetric CT by, e.g., forming helical paths in various directions and combining the resulting LBPs. The second approach in [112] computes 2D LBPs in three orthogonal planes and combines these.

In conclusion, we propose to use texture measures such as LBP for quantitative analysis of pulmonary emphysema in CT images of the lung. ROI classification experiments showed good classification performance, with an accuracy of 95.2%, and quantitative measures of emphysema derived by fusing posterior probabilities achieved high correlation with PFT, up to $|r| = 0.79$ ($p < 10^{-4}$). Overall, LBP seem to perform slightly better than a rotation invariant GFB, although the difference was not significant in our experiments. MCP_{NT} correlated significantly better with pulmonary function than the most common standard CT measure, RA, which suggests that texture based measures may be better indicators of the degree of emphysema. In addition, LBP seem to pick up certain micro-structures that are more frequent in smokers, including smokers who still have good lung function, than in people who never smoked. This structural information improves discrimination in our experiments and may also improve sensitivity to early changes in lung tissue integrity.

Chapter 3

Data-Driven Quantification of COPD

This chapter is based on the manuscript "Learning COPD Sensitive Filters in Pulmonary CT," by L. Sørensen, P. Lo, H. Ashraf, J. Sporring, M. Nielsen, and M. de Bruijne, published in Proc. *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 5761, pp. 699–706, 2009.

Abstract This chapter presents a fully automatic, data-driven approach for texture-based quantitative analysis of chronic obstructive pulmonary disease (COPD) in pulmonary computed tomography (CT) images. The approach uses supervised learning where the class labels are, in contrast to previous work, based on pulmonary function tests (PFTs) instead of on manually annotated regions of interest (ROIs). A quantitative measure of COPD is obtained by fusing probabilities computed in ROIs within the lung fields where the individual ROI probabilities are computed using a k nearest neighbor (kNN) classifier trained on a set of randomly sampled ROIs from the training CT images. The sampled ROIs are labeled using PFT data. The distance between two ROIs in the kNN classifier is computed as the textural dissimilarity between the ROIs, where the ROI texture is described by histograms of filter responses from a multi-scale, rotation invariant Gaussian filter bank. On 296 images taken from a lung cancer screening trial, the proposed measure was significantly better at discriminating between subjects with and without COPD than were the two most common computerized quantitative measures of COPD in the literature, namely, relative area of emphysema (RA) and percentile density (PD). The proposed measure achieved an AUC of 0.823 and correlated significantly with lung function whereas PD, the best performing alternative, achieved an AUC of 0.589 and did not correlate significantly with lung function. The proposed measure was also shown to be more reproducible and less influenced by inspiration level compared to RA and PD.

3.1 Introduction

Current quantitative measures of chronic obstructive pulmonary disease (COPD) are limited in several ways. The gold standard for diagnosis of COPD is pulmonary function tests (PFTs) [77]. These non-invasive measurements are cheap and fast to acquire but are limited by insensitivity to early stages of COPD [40] and lack of reproducibility [22]. Visual and computerized assessment in computed tomography (CT) imaging has emerged as an alternative that directly can measure the two components of COPD, namely, chronic bronchitis and emphysema. However, it is difficult to visually assess disease severity and progression. Moreover, visual assessment is subjective, time-consuming, and suffers from intra-observer and inter-observer variability [6, 58]. The most widely used computerized measures, also referred to as densitometry or quantitative CT, are the relative area of CT attenuation values below a certain threshold (RA) [58] and percentile density (PD) [36]. These measures consider only emphysema and treat each parenchyma voxel in the CT image independently, thereby disregarding potentially valuable information such as spatial relations between voxels and patterns at larger scales. The measures are also restricted to a single threshold parameter, which makes them sensitive to scanner calibration and noise in the CT images.

Supervised texture classification in CT, where a classifier is trained on manually annotated regions of interest (ROIs) [12, 65, 75, 87, 93, 102, 111], uses much more of the information available in the CT images compared to the densitometric measures, and the output of a trained classifier can be used for COPD quantification, e.g., by fusing individual ROI posterior probabilities [55, 65, 93]. However, this approach requires labeled data, which is usually acquired by manual annotation done by human experts. Manual annotation suffers from the same limitations as visual assessment of emphysema in CT images [6, 58], moreover, it is hard to accurately outline regions of emphysema since the appearance of the disease patterns can be subtle and diffuse, especially at early stages of COPD. Further, analysis is limited to current knowledge and experience of the experts, and there can be a bias towards typical cases in the annotated data set. As a consequence, unknown or less typical patterns that are a characteristic part of COPD may not captured by the trained classifier, and important discriminative information may be disregarded.

In this chapter, we propose a completely data-driven approach to texture-based analysis of COPD in pulmonary CT images. The main idea is to utilize meta-data that is connected with the CT images to acquire the labels. Hereby, no human intervention is required, and all the above mentioned limitations are handled. Instead, a fully data-driven, and thereby objective, CT image texture-based measure is obtained that can easily be applied to analyze large data sets. Other studies using labels acquired from meta-data, with different features and classification setup, have been published in other areas of medical image analysis as well, including assessment of structural changes of the breast tissue in digital mammography [78] and detection of tuberculosis in chest radiographs [3].

The proposed approach relies on supervised texture-based classification of ROIs and fusion of individual ROI posterior probabilities similar to [55, 65, 93], but with

ROIs and labels obtained in the following way: each CT image is assigned a global label according to PFTs of the scanned subject that are acquired at the same time as the CT image, and ROIs are sampled at random from within the lung fields and labeled with the global label of the CT image. In principle, other meta-data associated with the subject from which the CT image is acquired, such as genetic information and biomarkers blood samples, could be used when labeling. In this chapter, PFTs are used, which are the current gold standard for diagnosis of COPD [77]. The obtained texture-based measure is a probability of a subject suffering from COPD based on the evidence in the CT image, and this number reflects COPD severity in two ways. It measures the number of ROIs that show signs of COPD, i.e., how much of the lungs are affected, as well as the individual ROI COPD probabilities, i.e., the confidence about abnormality in individual ROIs.

The performance of the obtained texture-based measure is compared to the performance of the common densitometric measures RA and PD on a two-class classification problem defined using PFTs, i.e., diagnosis of COPD. The stability of the approach w.r.t. the randomly sampled ROIs is also inspected, and the reproducibility of the approach, as well as its robustness to inspiration level – a major source of variability in CT images, is further evaluated and compared to that of RA and PD.

3.2 Methods

The proposed quantitative measure for COPD relies on texture-based classification of CT ROIs. The ROI classification is done with a k nearest neighbor (kNN) classifier using dissimilarity between sets of filter response histograms as distance, and the histograms are based on the filter responses from a rotation invariant, multi-scale Gaussian filter bank [97]. A quantitative measure of the severity of COPD is obtained by fusing the individual ROI posterior probabilities into one posterior probability [55]. This approach has previously been successfully applied on another CT data set using manually labeled ROIs for training [93]. In [93], the same histogram estimation technique was used with two-dimensional versions of a subset of the filters considered in this chapter, and a quantitative measure of severity was also obtained by fusing ROI posteriors as classified by kNN, however, the ROI posteriors were estimated using prototype distances.

A segmentation of the lung fields is used in order to steer the sampling of ROIs as well as to decide which voxels contribute to the filter response histograms, and Section 3.2.1 describes how this segmentation is obtained. The filter response histograms, or texture descriptors, are described in Section 3.2.3, the ROI classification scheme is described in Section 3.2.4, and the posterior probability fusion is described in Section 3.2.5.

3.2.1 Segmentation of the lung fields

The lung fields are segmented in CT image I using a region growing algorithm, which assumes that lung parenchyma is below -400 Hounsfield units (HU). The algorithm

automatically detects part of the trachea by searching for a dark cylindrical structure in the top of the image, and the detected trachea is subsequently used to segment the left and right main bronchi. The segmented left and right main bronchi are then used to initiate two region growing procedures that segment the left and right lung field. The final segmented lung fields, $s(I)$, are obtained after a post processing step, where erroneously included regions belonging to the esophagus are removed by looking for tube-like structures between the segmented left and right lung fields. This is the same lung segmentation algorithm as is used in [53]. $s(I)$ excludes the trachea, the main bronchi, and any structures with CT intensity above -400 HU, which includes part of the vessels, the fissures, and the airway walls.

3.2.2 Sampling of ROIs

N_r, possibly overlapping, cubic ROIs are sampled at random from within the lung fields of CT image I according to segmentation $s(I)$, and these ROIs represent that image. Only ROIs with centers inside the segmentation are allowed, but parts of an ROI can still be outside the segmentation. These parts are disregarded in the subsequent analysis.

3.2.3 Texture descriptors

In this chapter, the textural information in a CT image is captured by measuring various texture features in randomly sampled ROIs from that image, and a filtering approach is used for this purpose. A filter bank comprising a total of eight rotational invariant filters based on the Gaussian function and combinations of derivatives of the Gaussian is applied at multiple scales, giving rise to a large number of filtered versions of the CT image. The ROIs in the image are represented by histograms of the filter responses, one for each of the applied filters, and classification is done based on this representation. Steps for obtaining the filter response histograms are given as follows:

Filters

Eight different measures of local image structure are used as base filters and these are: The Gaussian function

$$G(\mathbf{x};\sigma) = \frac{1}{(2\pi^{1/2}\sigma)^3}\exp\left(-\frac{||\mathbf{x}||_2^2}{2\sigma^2}\right) \tag{3.1}$$

where σ is the standard deviation, or scale, and $\mathbf{x} = [x, y, z]^T$ is a voxel; the three eigenvalues of the Hessian matrix

$$\lambda_i(\mathbf{x};\sigma), \quad i = 1, 2, 3, \quad |\lambda_1| \geq |\lambda_2| \geq |\lambda_3|; \tag{3.2}$$

gradient magnitude

$$||\nabla G(\mathbf{x};\sigma)||_2 = \sqrt{I_{x,\sigma}^2 + I_{y,\sigma}^2 + I_{z,\sigma}^2} \tag{3.3}$$

where $I_{x,\sigma}$ denotes the partial first order derivative of image I w.r.t. x at scale σ; Laplacian of the Gaussian

$$\nabla^2 G(\mathbf{x};\sigma) = \lambda_1(\mathbf{x};\sigma) + \lambda_2(\mathbf{x};\sigma) + \lambda_3(\mathbf{x};\sigma); \tag{3.4}$$

Gaussian curvature

$$K(\mathbf{x};\sigma) = \lambda_1(\mathbf{x};\sigma)\lambda_2(\mathbf{x};\sigma)\lambda_3(\mathbf{x};\sigma); \tag{3.5}$$

and the Frobenius norm of the Hessian

$$||H(\mathbf{x};\sigma)||_F = \sqrt{\lambda_1(\mathbf{x};\sigma)^2 + \lambda_2(\mathbf{x};\sigma)^2 + \lambda_3(\mathbf{x};\sigma)^2}. \tag{3.6}$$

Since histograms are used, the ordering of the voxels is disregarded and a classifier can therefore not automatically learn combinations of features such as the Laplacian of the Gaussian from the individual eigenvalues. Combinations of the eigenvalues, i.e., (3.4), (3.5), and (3.6), are therefore explicitly used in the representation.

Normalized convolution

The filtering is done by normalized convolution [48] with a lung fields segmentation, obtained as described in Section 3.2.1, as binary mask. The equation for normalized convolution is given by

$$I_\sigma = \frac{(S(\mathbf{x})I(\mathbf{x})) * G(\mathbf{x};\sigma)}{S(\mathbf{x}) * G(\mathbf{x};\sigma)} \tag{3.7}$$

where $*$ denotes convolution and the segmentation $S = s(I)$ computed from image I is used as an indicator function, marking whether $\mathbf{x}$ is a lung parenchyma voxel or not. Derivatives are computed on the Gaussian filtered images using finite differences.

Histogram estimation

The filter responses are quantized into filter response histograms. The bin widths are derived using adaptive binning [60]. This technique locally adapts the histogram bin widths to the data set at hand such that each bin contains the same mass when computing the histogram of all data. Only voxels in the considered ROI that belong to a lung segmentation S are used, and the resulting histogram is normalized to sum to one.

The number of histogram bins N_b computed from N_s voxels is determined according to

$$w = \frac{3.49\sigma}{\sqrt[3]{N_s}}$$

where w is the bin width and σ is the standard deviation of the distribution of the filter responses [82]. Letting $3.49\sigma/w$ express the number of bins we get

$$N_b = \sqrt[3]{N_s}, \tag{3.8}$$

and this is the expression we will use to determine the number of histogram bins as a function of the number of voxels.

3.2.4 Classification

ROI classification is performed using the kNN classifier [16, 43] with summed histogram dissimilarity as distance

$$D(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N_f} L(f_i(\mathbf{x}), f_i(\mathbf{y})) \tag{3.9}$$

where N_f is the number of filter response histograms, $L(\cdot, \cdot)$ is a histogram dissimilarity measure, and $f_i(\mathbf{x}) \in \mathbb{R}^{N_b}$ is the i'th filter response histogram with N_b bins estimated from an ROI centered on $\mathbf{x}$.

Three histogram dissimilarity measures L are considered, the L1-norm, the L2-norm, and the earth movers distance (EMD) [80]. The L1-norm and L2-norm are instances of the p-norm

$$L_p(H, K) = ||H - K||_p = \left(\sum_{i=1}^{N_b} |H_i - K_i|^p \right)^{1/p} \tag{3.10}$$

with $p = 1$ or $p = 2$ and where $H, K \in \mathbb{R}^{N_b}$ are histograms each with N_b bins. The histograms used in this chapter are normalized to sum to one, and thus L_1 is equivalent to $1-$histogram intersection [80]. EMD can be computed using (3.10) with $p = 1$ on cumulative versions of H and K when H and K are one dimensional, have equal number of bins, and equal mass [51], which is the case in this chapter. This histogram dissimilarity measure will be denoted by L_{EMD}.

3.2.5 Posterior probabilities

Two levels of posterior probabilities are considered in this chapter, ROI probabilities and subject probabilities. Note that subject and CT image is used interchangeably in this chapter and termed I. The ROI probability is based on the common k nearest neighbor posterior probability estimate [24]

$$P(\omega_i|\mathbf{x}, I) = \frac{k_{\omega_i}(\mathbf{x})}{k}, \quad \mathbf{x} \in s(I) \tag{3.11}$$

where $k_{\omega_i}(\mathbf{x})$ is the number of nearest neighbors of the ROI centered on voxel $\mathbf{x}$, from lung segmentation $s(I)$, belonging to class ω_i out of a total of k nearest neighbors according to (3.9). The ROI posterior probabilities are combined into an overall subject posterior probability using a static fusion scheme, namely, the mean rule [55]

$$P(\omega_i|I) = \frac{1}{N_r} \sum_{j=1}^{N_r} P(\omega_i|\mathbf{x}_j, I) \tag{3.12}$$

where N_r is the number of ROIs that are considered. The average sample posterior probability (3.12) then provides a measure of the probability that a subject suffers from COPD, based on the CT image. This number reflects both the number of samples that show signs of COPD as well as the probability for the individual ROIs.

Table 3.1: Group characteristics and lung function measurements for the healthy and the COPD group in data set $\mathcal{A}$. The numbers reported are mean values, with standard deviation in parentheses and range in square brackets.

	Healthy	**COPD**
Age (years)	57 (5) [49-70]	57 (5) [49-69]
Sex (men/women)	105/39	32/120
Height (cm)	176 (8) [156-192]	170 (8) [150-190]
Weight (kg)	80 (13) [51-117]	69 (13) [40-112]
Pack years	33 (11) [10-88]	40 (18) [20-133]
Smoking status (former/current)	55/89	30/122
FEV_1 (L)	3.38 (0.64) [2.17-5.20]	1.92 (0.42) [0.85-2.78]
FEV_1%pred	111 (18) [81-152]	63 (11) [31-80]
FEV_1/FVC	0.76 (0.04) [0.70-0.87]	0.60 (0.08) [0.37-0.70]

3.3 Experiments and results

3.3.1 Data

Experiments are conducted using low-dose volumetric CT images acquired at full inspiration from current and former smokers enrolled in the Danish Lung Cancer Screening Trial (DLCST) [67] with the following scan parameters: tube voltage 120 kV, exposure 40 mAs, slice thickness 1 mm, and in-plane resolution ranging from 0.72 to 0.78 mm. The subjects were scanned at entry (baseline) and were then subsequently scanned annually (followup) for four consecutive years. Annual PFTs were also performed along with the CT images, including the forced expiratory volume in one second (FEV_1) and the forced vital capacity (FVC). Subjects were re-scanned after approximately three months in cases where non-calcified nodules with a diameter of 5 to 15 mm were detected.

We perform, experiments on two subsets of the DLCST database that we denote data set $\mathcal{A}$ and $\mathcal{B}$. These data sets are defined in the following way:

Data set $\mathcal{A}$: Two subject groups are defined using the Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria [77] that are based on FEV_1 and FVC, as well as FEV_1 corrected for age, sex, and height (FEV_1%pred): a healthy group (no COPD; $FEV_1/FVC \geq 0.7$) and a COPD group (GOLD stage II or higher; $FEV_1/FVC < 0.7$ and $FEV_1\%pred < 80\%$). We further enforce that the criteria should be fulfilled both at baseline and at first year followup in order to decrease the influence of PFT noise on the labels. The number of subjects in the groups are 144 healthy and 152 COPD subjects, and the baseline CT images of these subjects from the DLCST database are used. The characteristics of the two groups are reported

in Table 3.1. Each CT image is represented by N_r ROIs that are randomly sampled within the lung fields.

Data set $\mathcal{B}$: 50 CT image pairs from the DLCST database. Both images in a pair are from the same subject that has been re-scanned for a suspicious nodule and there is therefore a short time between the two scans. All pairs have less than 86 days between the acquisition dates, and the disease is not expected to progress far enough to induce visible differences in CT within this time interval. There is no overlap between the 50 CT image pairs used here and the baseline CT images in data set $\mathcal{A}$.

3.3.2 Training and parameter selection

There are several parameters to select in the proposed classification system and these are listed below together with the possible parameter values considered:

- ROI size $r \times r \times r$ with $r = \{21, 31, 41\}$ voxels;
- number of nearest neighbors in the kNN classifier $k = \{25, 35, 45\}$;
- histogram dissimilarity measure $L = \{L_1, L_2, L_{\text{EMD}}\}$;
- the different base filters $\{(3.1), (3.2), (3.3), (3.4), (3.5), (3.6)\}$ at scales $\sigma = \{0.6(\sqrt{2})^i\}_{i=0,\ldots,6}$ mm.

The best combination of r, L, and k is learned from the training set, $\mathcal{T}$, and sequential forward feature selection (SFS) [43] is used for determining the optimal histogram subset for each combination. Together with the original intensity histogram, a total of $N_f = 57$ histograms are considered in the SFS. The CT images in the training set, $\mathcal{T} = \{I_i\}$, are divided into a prototype set $\mathcal{P}$ and a validation set $\mathcal{V}$ by randomly placing half the images of each group in each set. The classification system is trained by using the samples of $\mathcal{P}$ as prototypes in the kNN classifier and by choosing the histograms and parameter settings that minimize the classification error on $\mathcal{V}$. The number of ROIs sampled per subject, N_r, is fixed to 50, and the number of histogram bins is $N_b = r$ according to (3.8). The adaptive histogram binning is computed from the training set using a separate randomly sampled set of ROIs, where 10 ROIs are sampled per subject in the training set. kNN classification is performed using the approximate nearest neighbor (ANN) library [2] with the approximation error set to zero to turn off the approximation part of the algorithm.

3.3.3 Evaluation

The performance of the classification system is estimated using 3-fold cross-validation. The system is trained in each fold as described above, and the test set is classified using the best performing kNN classifier with the samples of the complete training set, $\mathcal{T} = \mathcal{P} \cup \mathcal{V}$, as prototypes. The results are evaluated by receiver operating characteristic (ROC) analysis and by correlation with FEV_1%pred using Pearson's correlation coefficient.

Table 3.2: COPD diagnosis and quantification results. AUCs from the ROC analysis with p-values for difference in AUC with kNN according to a DeLong, DeLong, and Clarke-Pearson's test [17] shown in parenthesis. Correlation with FEV_1%pred according to Pearson's correlation coefficient with p-values of in parenthesis.

Measure	**AUC**	**Correlation with FEV_1%pred**
kNN	0.823 (-)	0.48 ($p < 10^{-4}$)
RA_{950}	0.585 ($p < 10^{-4}$)	-0.10 ($p = 0.073$)
PD_{15}	0.589 ($p < 10^{-4}$)	0.10 ($p = 0.086$)

We compare the obtained results to the densitometric measures RA and PD. The densitometric measures are computed from the entire lung fields according to a lung segmentation $s(I)$. RA corresponds to the amount of voxels below a given HU threshold relative to the total amount of voxels within the lung fields segmentation, where the used threshold is close to that of air, -1000 HU, [58]. This measure is sometimes referred to as emphysema index or density mask. PD is derived from the CT attenuation histogram as the HU value at a certain percentile, usually the 15^{th} [36]. In this chapter, a HU threshold of -950 is used in RA, denoted RA_{950}, and the 15^{th} percentile is used in PD, denoted PD_{15}. The proposed measure is denoted kNN in the experiments.

3.3.4 COPD diagnosis and quantification

The whole learning framework is applied to data set $\mathcal{A}$ for COPD diagnosis and quantification using the resulting quantitative measure. The results of the experiment are shown in Fig. 3.1 and in Table 3.2. The proposed texture-based approach, achieving an AUC of 0.823, is significantly better at discriminating between CT images from healthy subjects and COPD subjects than are the densitometric measures PD_{15} and RA_{950}, $p < 10^{-4}$. kNN is significantly correlated with FEV_1%pred whereas PD_{15} and RA_{950} are not. All three evaluated measures are capable of separating the two subject groups, $p < 0.05$, according to a Wilcoxon rank sum test.

Note that the densitometric measures are computed from the full lung fields, and they are therefore based on more information than are the proposed texture-based measure, which is computed from 50 randomly sampled ROIs. The performance of PD_{15} and RA_{950} when computed only from the same 50 ROIs as used in kNN, is slightly worse than when computed from the entire lung fields with AUC = 0.584 and AUC = 0.577, respectively.

3.3.5 Stability of proposed measure

To inspect whether $N_r = 50$ is a sufficient number of samples in order to capture the characteristics in data set $\mathcal{A}$ related to healthy subjects and COPD subjects, we repeated the whole learning procedure ten times. In each repeated procedure, the

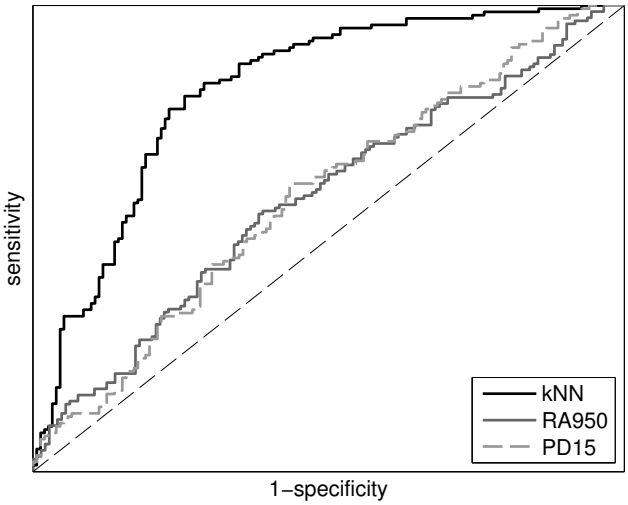


Figure 3.1: ROC curves from the experiment. The curve for kNN is based on (3.12).

Table 3.3: Number of times a certain parameter is selected out of 30 possible (10 repeated 3-fold cross-validations) in the ten repeated experiments.

k**NN**		**ROI size**		**histogram dissimilarity**	
$k = 25$	6	$r = 21$	0	$L = L_1$	15
$k = 35$	7	$r = 31$	0	$L = L_2$	12
$k = 45$	17	$r = 41$	30	$L = \text{EMD}$	3

same training, prototype, validation, and test data splits were used, but each time with different randomly sampled ROIs. Fig. 3.2 shows the resulting ROC curves and the AUCs are reported in the legend in the figure. The standard deviation on the AUCs is 0.015. The AUCs are rather similar, and they are much larger than the AUCs of the densitometric measures.

The selected parameters in the ten 3-fold cross-validations are reported in Table 3.3. The tendency is large k in the kNN classifier, large ROI size r, and L1-norm or L2-norm as histogram dissimilarity measure. Fig. 3.3 reports the number of times individual filters are selected and Table 3.4 reports the most commonly occurring filter subsets of sizes two and three. The gradient magnitude, $||\nabla G(\mathbf{x}; \sigma)||_2$, is by far the most often selected base-filter. The filter is selected 21 times at $\sigma = 4.8$ mm and 17 times at $\sigma = 2.4$ mm, out of 30 possible times. Further, the filter is member of all the most commonly selected subsets of sizes two and three, sometimes at multiple scales. The remaining base-filters are selected at least five times at a certain scale,

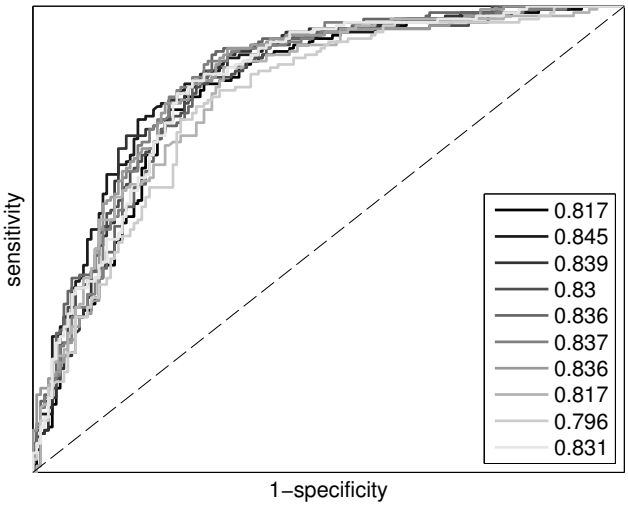


Figure 3.2: ROC curves for kNN in ten repeated experiments with different random ROI samplings on the same subject data splits. The legend shows the AUC of each experiment.

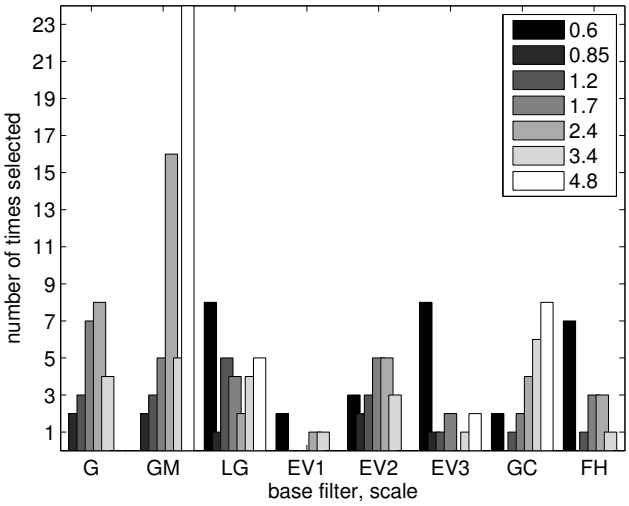


Figure 3.3: Number of times a certain filter is selected out of 30 possible (10 repeated 3-fold cross-validations) in the ten repeated experiments, grouped by base filter. Scales are ranging from black: 0.6 mm to white: 4.8 mm. The abbreviations are: Gaussian (G), gradient magnitude (GM), Laplacian of the Gaussian (LG), first, second, and third eigen value of the Hessian Matrix (EV1), (EV2), and (EV3), Gaussian curvature (GC), and Frobenius norm of the Hessian (FH).

Table 3.4: Filter subsets of sizes two and three that were most often selected in the SFS.

filter subset	occurrences
$\{\|\|\nabla G(\mathbf{x};2.4)\|\|_2, \|\|\nabla G(\mathbf{x};4.8)\|\|_2\}$	14
$\{\lambda_3(\mathbf{x};0.6), \|\|\nabla G(\mathbf{x};4.8)\|\|_2\}$	9
$\{\lambda_3(\mathbf{x};0.6), \|\|\nabla G(\mathbf{x};2.4)\|\|_2\}$	8
$\{\nabla^2 G(\mathbf{x};0.6), \|\|\nabla G(\mathbf{x};2.4)\|\|_2\}$	7
$\{K(\mathbf{x};4.8), \|\|\nabla G(\mathbf{x};2.4)\|\|_2\}$	7
$\{\nabla^2 G(\mathbf{x};0.6), \|\|\nabla G(\mathbf{x};4.8)\|\|_2\}$	6
$\{K(\mathbf{x};4.8), \|\|\nabla G(\mathbf{x};4.8)\|\|_2\}$	6
$\{\|\|H(\mathbf{x};0.6)\|\|_F, \|\|\nabla G(\mathbf{x};4.8)\|\|_2\}$	6
$\{K(\mathbf{x};3.4), \|\|\nabla G(\mathbf{x};4.8)\|\|_2\}$	6
$\{\nabla^2 G(\mathbf{x};0.6), \|\|\nabla G(\mathbf{x};2.4)\|\|_2, \|\|\nabla G(\mathbf{x};4.8)\|\|_2\}$	6
$\{\lambda_3(\mathbf{x};0.6), \|\|\nabla G(\mathbf{x};2.4)\|\|_2, \|\|\nabla G(\mathbf{x};4.8)\|\|_2\}$	6
$\{K(\mathbf{x};4.8), \|\|\nabla G(\mathbf{x};2.4)\|\|_2, \|\|\nabla G(\mathbf{x};4.8)\|\|_2\}$	6

except for the absolute largest and second largest eigenvalues of the Hessian matrix, $\lambda_1(\mathbf{x};\sigma)$ and $\lambda_2(\mathbf{x};\sigma)$, which are rarely selected. Frequently occurring subsets are the Laplacian of the Gaussian at a small scale together with the gradient magnitude at a large scale, and the absolute smallest eigenvalue at a small scale together with the gradient magnitude at a large scale, see Table 3.4. SFS typically selects $5-7$ histograms out of the 57 possible histograms.

3.3.6 Reproducibility and robustness to inspiration level

The reproducibility of the proposed measure as well as the robustness to inspiration level is evaluated and compared to the densitometric measures on data set $\mathcal{B}$. The trained kNN classifiers from the three folds in the experiment in Section 3.3.4 are used to represent the proposed measure.

The reproducibility of a measure is evaluated by the correlation, measured using Spearman's rank correlation, between the vector of measurements obtained from the first image in the 50 pairs, $\mathbf{m}_1$, of data set $\mathcal{B}$ and the vector of measurements on the second image in the pairs, $\mathbf{m}_2$. The results are reported in the second column of Table 3.5. kNN is more reproducible than RA_{950} and PD_{15} in all three folds.

The main source of variability between two CT images from the same subject,

Table 3.5: Measures of reproducibility and robustness to inspiration level, both using Spearman's rank correlation $\rho(\cdot,\cdot)$. p-values of the correlations are shown in parentheses.

Measure	$\rho(\mathbf{m}_1, \mathbf{m}_2)$	$\rho(\mathbf{m}_2 - \mathbf{m}_1, \mathbf{LV_{rd}})$
kNN, fold 1	0.90 ($p < 10^{-4}$)	-0.40 ($p = 0.004$)
kNN, fold 2	0.88 ($p < 10^{-4}$)	-0.12 ($p = 0.401$)
kNN, fold 3	0.88 ($p < 10^{-4}$)	-0.35 ($p = 0.012$)
RA_{950}	0.82 ($p < 10^{-4}$)	0.83 ($p < 10^{-4}$)
PD_{15}	0.81 ($p < 10^{-4}$)	-0.82 ($p < 10^{-4}$)

with a short time interval between acquisition dates, is expected to be the inspiration level. However, other sources of variability, such as scanner drift and different subject orientations during scanning also play a role. We use the lung volume (LV) as an indicator of the inspiration level. The sensitivity to inspiration level is evaluated by correlating, using Spearman's rank correlation, signed measurement difference, $\mathbf{m}_2 - \mathbf{m}_1$, with relative signed LV difference, computed as the difference divided by the average and denoted by $\mathrm{LV}_{\mathrm{rd}}$. The results are reported in the third column of Table 3.5. Differences in both the densitometric measures are significantly correlated with LV difference whereas differences in one out of three kNN based measures is not. The proposed measure is therefore, for certain trained kNN classifiers, less sensitive to inspiration level than are the densitometric measures.

3.4 Discussion and conclusion

The conducted experiments show that it is possible to train a texture-based classifier to recognize COPD in pulmonary CT images using supervised learning techniques in a fully automatic, data-driven approach without any human intervention. Hereby, all the limitations associated with manual labeling are avoided. The meta-data driven labeling of ROIs, in this chapter using PFTs, however, has other potential problems. The disease patterns may be localized only in parts of the CT images in subjects with COPD. For instance, paraseptal emphysema is located in the periphery of the lung, centrilobular emphysema is predominantly in the upper lobes, while panlobular emphysema is predominantly in the lower lobes [108]. Randomly sampled ROIs from COPD subjects will therefore likely contain both diseased and healthy tissue where the healthy tissue ROIs still receive the label COPD. The reverse may also be the case in healthy subjects but is expected to be less prominent. The classes in this weakly labeled data set are therefore expected to overlap more compared to classes in manually labeled data where experts have annotated relatively clear examples of the different classes, and this poses a challenging classification problem.

Intensity can be directly related to emphysema in CT since emphysematous regions have lower attenuation than do healthy regions due to loss of lung tissue. Orig-

inal and smoothed intensities, $G(\mathbf{x};\sigma)$, may, therefore, be considered as important features when discriminating between lung tissue in healthy subjects and in COPD subjects. This is also supported by the results of the stability experiment in Section 3.3.5 where $G(\mathbf{x};\sigma)$ is selected 16 out of 30 possible times in the SFS procedure. However, the original intensity, which is what RA and PD rely on, is never selected. Some filters capturing structural information are also selected often, $||\nabla G(\mathbf{x};\sigma)||_2$ is selected 27 times, $\nabla^2 G(\mathbf{x};\sigma)$ is selected 21 times, and $K(\mathbf{x};\sigma)$ is selected 18 times. Consequently, the resulting classifiers use smoothed intensity information in conjunction with first and second order derivative information, which makes the proposed measure for COPD very different from the standard densitometric measures RA and PD. The use of information at larger scales, the use of texture information, and the fact that entire histograms are used instead of one measure summarizing each histogram, may explain the improved performance compared to RA and PD.

The general tendency for the filters capturing structural information that are selected in the SFS procedure is large scale edges, $||G(\mathbf{x};4.8)||_2$ and/or $||G(\mathbf{x};2.4)||_2$, in conjunction with small scale blobs, $\nabla^2 G(\mathbf{x};0.6)$, or large scale blobs, $K(\mathbf{x};4.8)$, see Table 3.4. These results share both similarities and dissimilarities with a previous study using a similar classification setup with a subset of the filters used in this chapter on a different data set, but with the important difference that manually annotated ROIs were used [93]. $||G(\mathbf{x};\sigma)||_2$ at a large scale is almost always selected, both in the present chapter and in [93]. On the contrary, $G(\mathbf{x};\sigma)$ at a small scale, i.e., intensity information, is also frequently selected in [93] whereas it is less frequently selected in the present chapter. This may be explained by the weak labeling of data causing a large class overlap.

The gender distribution in the two groups is skewed, see Table 3.1. The ROC-analysis is therefore repeated for each gender separately to inspect whether the gender skewness has influenced the results. For males we get the following AUCs: 0.879, 0.770, and 0.781, for kNN, RA_{950}, and PD_{15}, respectively, and for females the AUCs are: 0.808, 0.641, and 0.636. For both genders, the AUC of kNN is larger compared to the densitometric measures, and it is significantly larger in all but one case. For males, the kNN AUC is not significantly larger than the PD_{15} AUC, $p = 0.11$.

PFTs are insensitive to early stages of COPD [40], lack reproducibility [22], and can be affected by other factors limiting the airflow in the airways than those associated with COPD. Despite these limitations, PFTs were used to obtain labels in this chapter assuming that it was possible to learn, using supervised learning, the textural COPD patterns in CT that are related to the part of the disease that correlates with PFTs. PFTs are also the current gold standard for COPD diagnosis [77]. A two-class problem was defined by the two subject groups, healthy (no COPD according to the GOLD criteria [77]), and COPD (GOLD stage II or higher according to the GOLD criteria [77]). However, other possibilities exist, both on the type of problem to consider and on the type of meta-data to use for group definitions. One possibility would be to consider several or all of the four GOLD stages [77] as separate groups, which is similar in spirit to [59], for assessing GOLD stage or COPD severity. However, regression may be more suitable for this purpose. The proposed approach

could also be used to gain a better understanding of which textural patterns in the CT images that are related to specific genes or markers from blood samples by using genetic information or blood biomarkers to define groups and apply the whole learning framework. This may be expanded to further analyze how these patterns evolve over time in longitudinal data.

Classifiers were trained at the ROI level without including knowledge about the subsequent fusion of the ROI posterior probabilities using (3.12). The rationale is that we would like to capture the local texture information and use this for quantification. Although the proposed approach works well as illustrated in the results, it remains an open research question whether training locally followed by posterior fusion is the best approach when the final goal is quantification at CT image level. An alternative approach would be to take CT image level information into account during training, e.g., by adapting the objective function for SFS to use (3.12) instead of (3.11).

COPD comprises two main components, small airway disease, or chronic bronchitis, and emphysema [77]. The proposed approach measures parenchymal texture and therefore mainly targets emphysema. However, small airway disease is included to some extent since the lung fields segmentation includes the small airways and since the labels are obtained from PFTs, which are affected by both components. The airway information could be targeted more explicitly, e.g., by combining the output of the proposed approach with measurements computed directly on the segmented airway tree, e.g., [71], which may provide a more accurate measure of COPD.

In conclusion, we have proposed a fully automatic, data-driven approach for texture-based quantitative analysis of COPD in pulmonary CT. The obtained texture-based measure demonstrates superior performance in discriminating between subjects with and without COPD compared to the common densitometric measures RA and PD, with an AUC of 0.823 compared to 0.585 and 0.589, respectively. The texture-based measure also correlates significantly with FEV_1%pred with a correlation coefficient of $r = 0.48$ compared to the non-significant correlations of the two densitometric measures. Further, the proposed approach is more reproducible and is less sensitive to inspiration level – a major source of variability in computerized quantitative CT measures, compared to the densitometric measures. Since the approach does not rely on labels annotated by human experts, the resulting CT image-based measure is objective and can easily be applied to analyze large data sets. Overall, the proposed approach results in a robust and objective measure that can facilitate better computerized quantification of COPD in pulmonary CT.

Chapter 4

Dissimilarity-Based Classification

This chapter is based on the manuscript "Dissimilarity Representations in Lung Parenchyma Classification," by L. Sørensen and M. de Bruijne, published in Proc. *SPIE Medical Imaging: Computer-Aided Diagnosis*, SPIE Press, vol. 7260, pp. 72602Z1-72602Z12, 2009.

Abstract A good problem representation is important for a pattern recognition system to be successful. The traditional approach to statistical pattern recognition is feature representation. More specifically, objects are represented by a number of features in a feature vector space, and classifiers are built in this representation. This is also the general trend in lung parenchyma classification in computed tomography (CT) images, where the features often are measures on feature histograms. Instead, we propose to build normal density based classifiers in dissimilarity representations for lung parenchyma classification. This allows for the classifiers to work on dissimilarities between objects, which might be a more natural way of representing lung parenchyma. In this context, dissimilarity is defined between CT regions of interest (ROI)s. ROIs are represented by their CT attenuation histogram and ROI dissimilarity is defined as a histogram dissimilarity measure between the attenuation histograms. In this setting, the full histograms are utilized according to the chosen histogram dissimilarity measure.
We apply this idea to classification of different emphysema patterns as well as normal, healthy tissue. Two dissimilarity representation approaches as well as different histogram dissimilarity measures are considered. The approaches are evaluated on a set of 168 CT ROIs using normal density based classifiers all showing good performance. Compared to using histogram dissimilarity directly as distance in a k nearest neighbor classifier, which achieves a classification accuracy of 92.9%, the best dissimilarity representation based classifier is significantly better with a classification accuracy of 97.0% ($p = 0.046$).

4.1 Introduction

The traditional approach to statistical pattern recognition is feature representation. More specifically, objects are represented by a number of features in a feature vector space, and classifiers are built in this representation [24]. This is also the general trend in lung parenchyma classification [12,65,87,102,111]. Duin *et al.* motivated the idea of basing classification directly on distances between objects, thereby completely avoiding features [25]. Instead of focusing on finding good features for describing objects, the focus is moved to finding good dissimilarity measures for comparing objects. Dissimilarity representations may be preferable to the traditional feature representation approach, e.g., when there is not enough expert knowledge available to define proper features or when data is high dimensional [68].

Working in a dissimilarity representation of objects, a k nearest neighbor (kNN) classifier [43], which is applied directly on distances between objects, is a natural and simple choice. However, there exist techniques that make it possible to use other classifiers such as normal density based classifiers on dissimilarity data [68]. The general idea is to represent data by a distance matrix containing pair-wise dissimilarities between objects, also called dissimilarity representation. From this representation, a feature space is derived in which traditional pattern recognition techniques then can be applied. Embedding of a Euclidean dissimilarity representation into a Euclidean space via classical scaling is one way of doing this [70]. A second approach is to treat the dissimilarity representation as a new data set with the rows being observations and the columns being dimensions in a dissimilarity space. Each dimension in this space measures the dissimilarity to a particular training prototype, and the set of prototypes is called the representation set [68]. A third approach that will not be considered further in this chapter, is embedding in a pseudo-Euclidean space in the case of a non-Euclidean dissimilarity representation [35, 70].

Compared to a density based classifier built in a dissimilarity space, kNN has high computational complexity and large storage requirements. In kNN, distances to all training set objects need to be computed when classifying novel patterns, and therefore the entire training set needs to be stored. In a dissimilarity space, a few objects can be selected from the training set as prototypes in the representation set, keeping the dimensionality low and only requiring storage of the representation set and the trained classifier. A kNN classifier makes the classification decision based only on a local neighborhood, i.e., the k closest prototypes, which makes it sensitive to noise. Density based classifiers in a dissimilarity representation are more global, in the sense that parameters of Gaussian functions are estimated off-line using all available dissimilarity training data while still working in a low dimensional dissimilarity space or embedding, which has a natural smoothing effect. Also, the classification is based on a weighted combination of the dissimilarities between the novel pattern and the prototypes. These weights are estimated using the entire training set and thus "essential" prototypes are given more weight in the classification decision. A density based classifier is therefore expected to achieve better generalization when dealing with a small and noisy data set, especially in cases of normal distributed classes.

Previously, we investigated the use of feature histograms for lung disease pattern classification in computed tomography (CT) using a histogram dissimilarity measure directly as distance in a kNN classifier, which showed promising results [92]. In the literature, measures of histograms, such as moments of filter response histograms and measures on co-occurrence matrices, are often used as features in a feature space when classifying lung disease patterns in CT [12, 65, 87, 102, 111]. Using only the first few moments of a histogram might discard valuable information. Instead, using the full histogram for classification may improve classification accuracy [60]. This chapter investigates the possible benefit of building classifiers in a histogram dissimilarity representation compared to using histogram dissimilarity directly as distance in a kNN classifier. In light of the previous discussions, we see several possible benefits of using a density based classifier trained in a histogram dissimilarity representation for lung parenchyma classification. To our knowledge, this has not been investigated before.

Pekalska *et al.* have applied dissimilarity representations in numerous standard data sets, including handwritten digits, polygons, road signs, and chromosome band profiles [68, 69]. Dissimilarity representations have also been used in various other pattern recognition applications. Trosset *et al.* used dissimilarity representations for discriminating patients with Alzheimer's disease from normal elderly subjects in magnetic resonance images. The dissimilarities were based on hippocampal dissimilarity obtained from image registration deformations [98]. In this chapter, we represent images by histograms and construct dissimilarity representations based on histogram dissimilarities, which is an approach also taken by other authors. Bruno *et al.* used a dissimilarity representation based on symmetrized Kullback-Leibler divergence between RGB histograms for image retrieval [10]. Paclik *et al.* investigated the use of dissimilarity representations in hyperspectral data classification using various histogram dissimilarity measures [63].

The specific application of this chapter is classification of emphysema subtype and normal tissue in regions of interest (ROI), based on the CT attenuation histogram. Emphysema is a major component of chronic obstructive pulmonary disease (COPD) and is characterized by gradual loss of lung tissue. COPD is a growing health problem worldwide. In the United States alone, it is the fourth leading cause of morbidity and mortality, and it is estimated to become the fifth most burdening disease worldwide by 2020 [77]. Methods for reliable classification of emphysema in lungs are therefore of interest, since they may form the basis for computer-aided diagnosis. CT imaging is gaining more and more attention as a diagnostic tool for COPD, and it is a sensitive method for diagnosing emphysema, assessing its severity, and determining its subtype. Both visual and quantitative CT assessment are closely correlated with the pathological extent of emphysema [84]. Emphysema is usually classified into three subtypes, or patterns, in CT [108], and the two of the three subtypes we focus on in this chapter are the following: centrilobular emphysema (CLE), defined as multiple small low-attenuation areas; and paraseptal emphysema (PSE), defined as multiple low-attenuation areas in a single layer along the pleura often surrounded by interlobular septa visible as thin white walls.

4.2 Methods

This section describes the methodology that we use. Section 4.2.1 briefly describes how the attenuation histograms are computed from the ROIs. Section 4.2.2 describes three different histogram dissimilarity measures used for comparing histograms. Section 4.2.3 describes two dissimilarity representation approaches: the dissimilarity space approach and an embedding approach based on classical scaling. Both are based on a distance matrix that in turn is based on a histogram dissimilarity measure. Finally, Section 4.2.4 describes two classifiers, a linear discriminant and a quadratic discriminant classifier, that both will be trained and tested in the dissimilarity representations.

4.2.1 Histogram estimation

We represent each ROI by its attenuation histogram. The histogram is estimated using non-linear binning by choosing the histogram bins such that the total distribution of the attenuation values in the training set is approximately uniform [60]. All histograms are normalized to sum to one.

4.2.2 Histogram dissimilarity measures

Three histogram dissimilarity measures L are considered: one based on histogram intersection (HI) [95], earth movers distance (EMD) [80], and the L_2-norm. HI is given by

$$HI(H,K) = \sum_{i=1}^{N_b} \min(H_i, K_i)$$

where $H \in \mathbb{R}^{N_b}$ and $K \in \mathbb{R}^{N_b}$ are histograms each with N_b bins. $HI(\cdot,\cdot)$ is a similarity measure, and a dissimilarity measure based on this can be obtained by

$$L_{HI}(H,K) = 1 - HI(H,K). \tag{4.1}$$

All histograms considered in this chapter sum to one, thus $L_{HI}(\cdot,\cdot) \in [0,1]$. EMD is given by

$$L_{EMD}(H,K) = \sum_{i=1}^{N_b}\sum_{j=1}^{N_b} C_{ij}F_{ij} \tag{4.2}$$

where $C \in \mathbb{R}^{N_b \times N_b}$ is a ground distance matrix and $F \in \mathbb{R}^{N_b \times N_b}$ is a flow matrix. The flow matrix contains the optimal flows obtained by solving the transportation problem of moving the mass of H such that it matches the mass of K. The L_2-norm is given by

$$L_2(H,K) = \sqrt{\sum_{i=1}^{N_b}(H_i - K_i)^2}. \tag{4.3}$$

4.2.3 Dissimilarity representations

Computing all pairwise dissimilarities L between the objects from the set $\mathcal{A} = \{a_1, \ldots, a_n\}$ and the set $\mathcal{B} = \{b_1, \ldots, b_m\}$ we obtain the $n \times m$ dissimilarity, or distance, matrix [68, 69]

$$D_L(\mathcal{A}, \mathcal{B}) = \begin{pmatrix} L(a_1, b_1) & \ldots & L(a_1, b_m) \\ \vdots & \ddots & \vdots \\ L(a_n, b_1) & \ldots & L(a_n, b_m) \end{pmatrix}. \tag{4.4}$$

Using (4.4) with either (4.1), (4.2), or (4.3) as histogram dissimilarity, we obtain three different distance matrix representations of the data $D_{L_{HI}}(\mathcal{A}, \mathcal{B})$, $D_{L_{EMD}}(\mathcal{A}, \mathcal{B})$, and $D_{L_2}(\mathcal{A}, \mathcal{B})$.

Dissimilarity space

One way to utilize the distance matrix (4.4) is by extracting a representation set of prototypes $\mathcal{R}$. Given a training set $\mathcal{T}$, this approach selects a set of objects $\mathcal{R} \subseteq \mathcal{T}$ from $\mathcal{T}$. All objects in $\mathcal{T}$ are represented in a dissimilarity space, where the i'th dimension corresponds to the dissimilarity with prototype $\mathcal{R}_i \in \mathcal{R}$, i.e., we compute $D_L(\mathcal{T}, \mathcal{R})$ [68]. Selecting a representation set is conceptually similar to selecting a limited number of prototypes for the kNN classifier. However, where the prototypes define the kNN classifier independently of the remaining training set, $\mathcal{R}$ defines a dissimilarity space in which the entire training set is represented and used to train a classifier. The final classifier is therefore expected to be less sensitive to the specific choice of prototypes.

There are different ways of choosing the representation set, e.g., random selection or feature selection methods, in this context searching for prototypes. For simplicity, we will only consider random prototype selection in this chapter. Random selection has previously been found to give reasonable results [70].

Embedding

Instead of selecting prototypes, another approach is to embed $D_L(\mathcal{T}, \mathcal{T})$ in a vector space and reduce the dimensionality of this space. Standard inner product based techniques can be applied in this space.

A $D_L(\mathcal{T}, \mathcal{T})$ based on an Euclidean dissimilarity measure L can be perfectly embedded in an Euclidean space by classical scaling, which is a distance preserving linear mapping [70]. It is based on the positive definite Gram matrix

$$G = -\frac{1}{2} J (D_L \odot D_L) J$$

where $\odot$ denotes entry-wise matrix multiplication and the centering matrix $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ where n is the number of training set objects and $\mathbf{1} = [1, \ldots, 1]^T \in \mathbb{R}^n$. G is factorized using an eigendecomposition

$$G = Q \Lambda Q^T$$

where Λ is a diagonal matrix containing eigenvalues ordered by descending magnitude and Q is a matrix containing the corresponding eigenvectors. For $k \leq n$ non-zero eigenvalues, a k-dimensional Euclidean embedding is then obtained by

$$E = Q_k \Lambda_k^{\frac{1}{2}} \tag{4.5}$$

where $Q_k \in \mathbb{R}^{n \times k}$ contains the first k leading eigenvectors and $\Lambda_k \in \mathbb{R}^{k \times k}$ contains the square roots of the corresponding eigenvalues.

When $D_L(\mathcal{T}, \mathcal{T})$ is based on a non-Euclidean dissimilarity measure, G is not positive definite and therefore has negative eigenvalues. In this case, an Euclidean embedding cannot be obtained using (4.5) since the computations rely on square roots of the eigenvalues. This problem can be addressed by considering only positive eigenvalues and corresponding eigenvectors in (4.5) [70].

Two of the histogram dissimilarity measures used in this chapter, (4.1) and (4.2), are non-Euclidean and one, (4.3), is Euclidean. When using Euclidean distance, i.e., (4.3), classical scaling recovers the original $n \times N_b$ data matrix from the $n \times n$ distance matrix up to location, reflection, and rotation.

4.2.4 Classifiers

Two classifiers are evaluated in the different dissimilarity representations: a linear discriminant classifier (LDC) and a quadratic discriminant classifier (QDC) [24, 43]. These classifiers have previously shown to perform well in dissimilarity spaces [69]. Both are density based classifiers using multivariate Gaussian functions to represent classes $\omega_i = \{\mu_i, \Sigma_i\}$

$$G_i(\mathbf{x}; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{N/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right)$$

where N is the dimensionality of the input space and $\mathbf{x} \in \mathbb{R}^N$ is a position in the input space. In LDC, equal class covariance matrices Σ are assumed resulting in the following linear discriminant function

$$g_i(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P(\omega_i) \tag{4.6}$$

where Σ and the class sample means μ_i are estimated in the dissimilarity representation obtained from $D_L(\mathcal{T}, \mathcal{T})$ and $P(\omega_i)$ is the class prior. In QDC, each class covariance matrix Σ_i is estimated separately resulting in the following quadratic discriminant function

$$g_i(\mathbf{x}) = -\frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) + \log P(\omega_i). \tag{4.7}$$

The density based classifiers assigns class ω_i to observation $\mathbf{x}$ according to the maximum discriminant function

$$\hat{g}(\mathbf{x}) = \arg\max_i g_i(\mathbf{x}). \tag{4.8}$$

4.3 Experiments and results

The data used for the experiments originates from a set of thin-slice CT images of the thorax. CT was performed using GE equipment (LightSpeed QX/i; GE Medical Systems, Milwaukee, WI, USA) with four detector rows, using the following parameters: In-plane resolution 0.78×0.78 mm, 1.25 mm slice thickness, tube voltage 140 kV, and tube current 200 milliampere (mA). The slices were reconstructed using a high spatial resolution (bone) algorithm. A population of 25 patients, 8 healthy non-smokers, 4 smokers without COPD, and 13 smokers diagnosed with moderate or severe COPD according to lung function tests [77] were scanned in the upper, middle, and lower lung, resulting in a total of 75 CT slices.

Visual assessment of the leading pattern, either NT, CLE, or PSE, in each of the 75 slices was done individually by an experienced chest radiologist and a CT experienced pulmonologist. 168 non-overlapping ROIs of size 31×31 pixels were annotated in the slices, representing the three classes: NT (59 observations), CLE (50 observations), and PSE (59 observations). The NT ROIs were annotated in the non-smokers and the CLE and PSE ROIs were annotated in the smokers, within the area(s) of the leading pattern.

Figure 4.1 shows an ROI from each of the three classes, together with the CT slices in which they were annotated, and Figure 4.2 shows the attenuation histograms of all 168 ROIs estimated using the non-linear binning principle described in Section 4.2.1.

4.3.1 Visualizing dissimilarity spaces

Three prototypes are selected at random, one from each class, and the resulting pair-wise dissimilarity spaces are inspected by plotting the dissimilarities between all observations and one prototype versus the dissimilarities between all observations and second prototype. The results can be seen in Figure 4.3. The class separation is already quite good using only two prototypes and it can be expected to be even better when using more than two prototypes. In some cases, there is a tendency to degenerate behavior of the resulting spaces, e.g., in Figure 4.3(f) where the PSE samples almost reside on a line in the two-dimensional dissimilarity space.

4.3.2 Visualizing embeddings

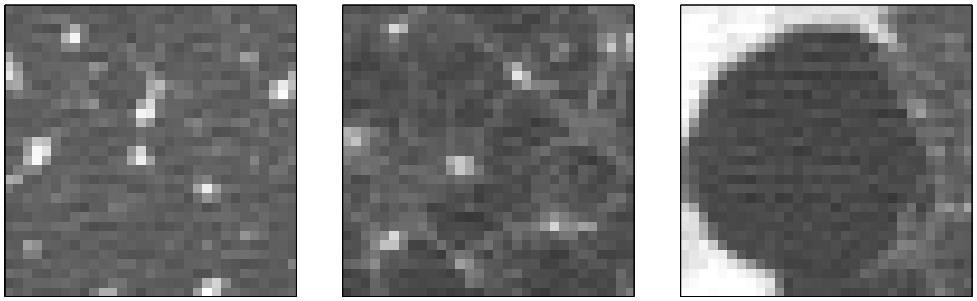
Figure 4.4 shows the eigenvalues derived in the embedding process for $D_{L_{HI}}$, $D_{L_{EMD}}$, and D_{L_2} on our data. As seen in Figure 4.4(a) and 4.4(b), the non-Euclidean property of L_{HI} and L_{EMD} is revealed by the presence of negative eigenvalues. The number of eigenvalues that are significantly different from zero is small in all three cases, showing that the intrinsic dimensionality of the three dissimilarity representations of the data is rather low.

Figure 4.5 shows two-dimensional embeddings obtained by using the two eigenvectors with the largest positive eigenvalues. The class separation is generally good in all three representations.

(a) CT slice with leading NT pattern.

(b) CT slice with leading CLE pattern.

(c) CT slice with leading PSE pattern.

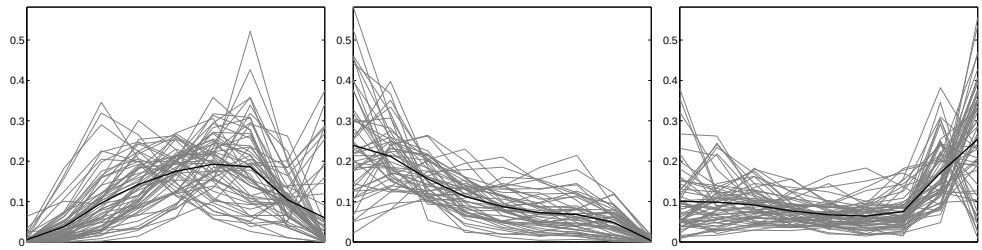


(d) NT ROI.

(e) CLE ROI.

(f) PSE ROI.

Figure 4.1: Examples slices and ROIs annotated in the same slices. The ROI in 4.1(d) is from 4.1(a) etc.

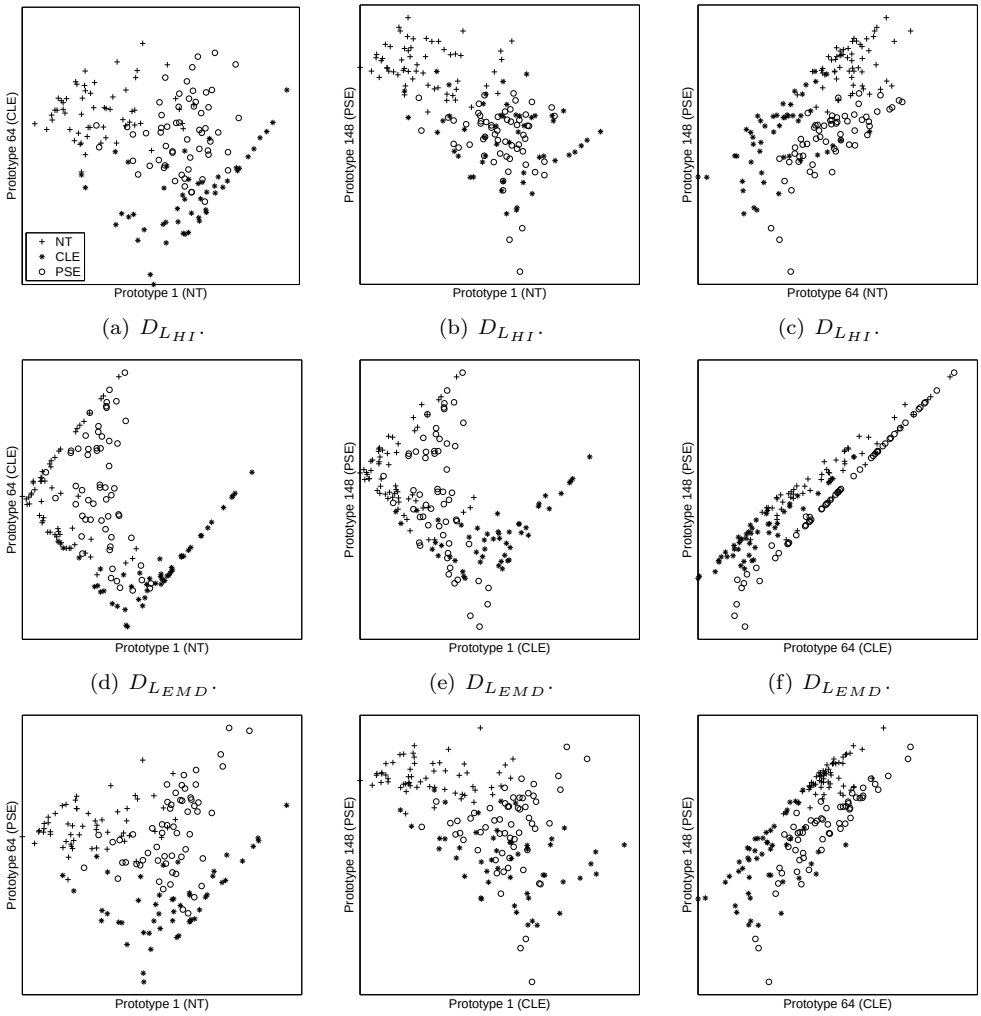


(a) NT.

(b) CLE.

(c) PSE.

Figure 4.2: Attenuation histograms estimated from the data. Individual histograms are shown in gray and the mean histogram is shown in black.

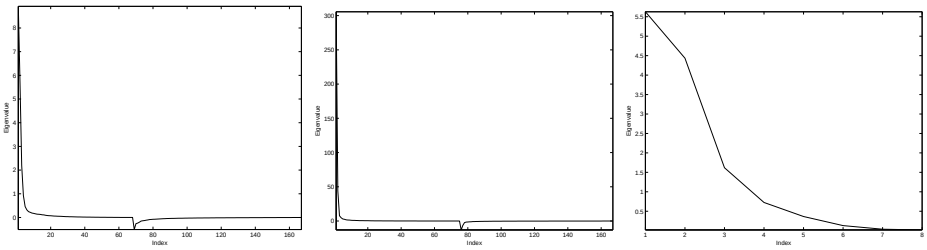


(a) $D_{L_{HI}}$. (b) $D_{L_{HI}}$. (c) $D_{L_{HI}}$.

(d) $D_{L_{EMD}}$. (e) $D_{L_{EMD}}$. (f) $D_{L_{EMD}}$.

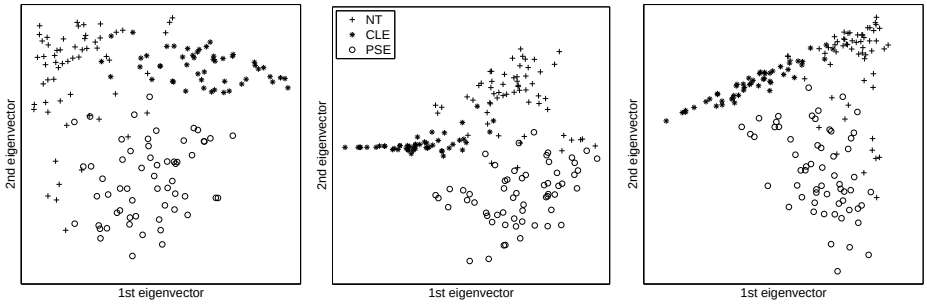
(g) D_{L_2}. (h) D_{L_2}. (i) D_{L_2}.

Figure 4.3: Examples of dissimilarity spaces obtained using representation sets with two random prototypes.



(a) $D_{L_{HI}}$. (b) $D_{L_{EMD}}$. (c) D_{L_2}.

Figure 4.4: Eigenvalues derived in the embedding process sorted by absolute value. In 4.4(a) and 4.4(b) the eigenvalues are divided in a positive and a negative part.



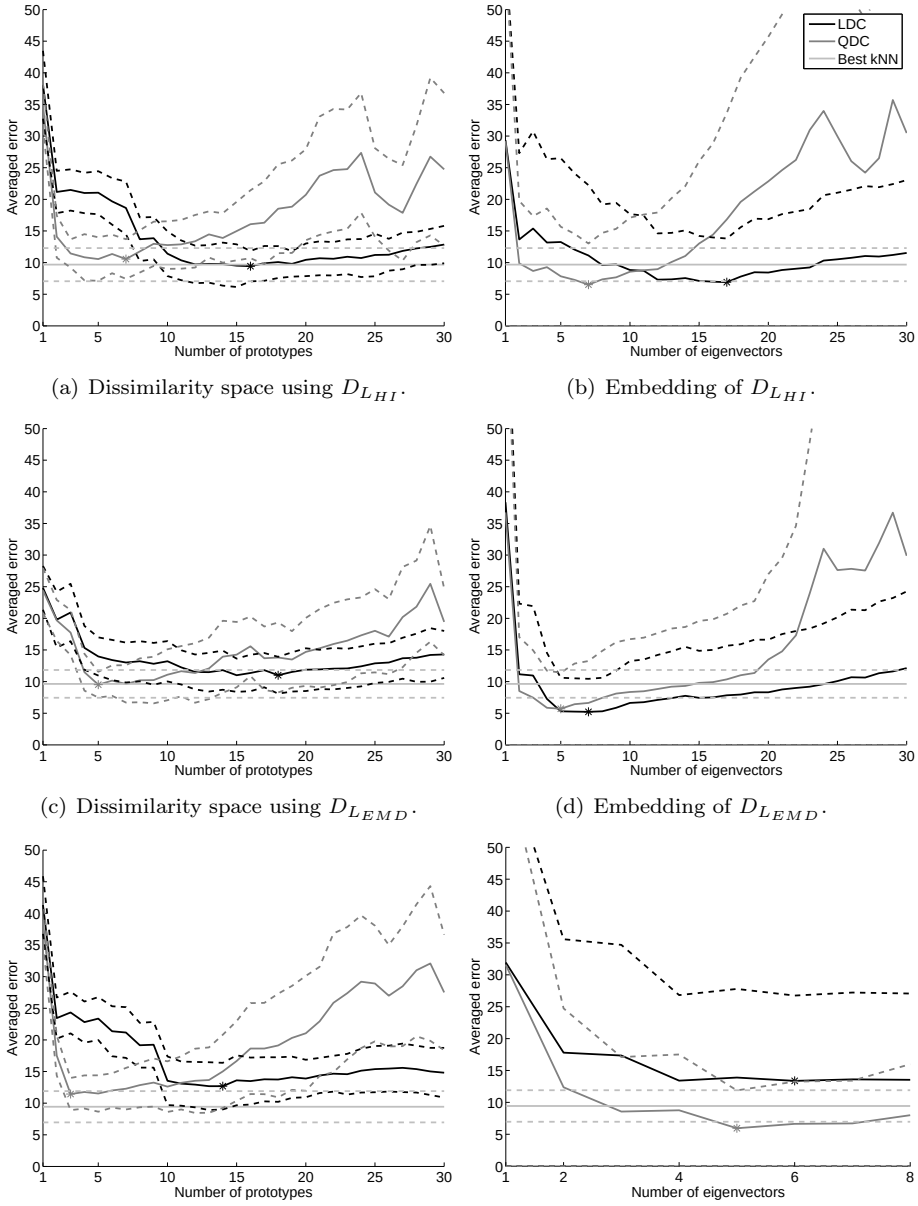
(a) $D_{L_{HI}}$. (b) $D_{L_{EMD}}$. (c) D_{L_2}.

Figure 4.5: Two-dimensional embedding of D_L using the two eigenvectors with the largest positive eigenvalue.

4.3.3 Classifier stability

We use feature curves for inspecting the stability of the dissimilarity representation based classifiers as a function of the number of dimensions in the representation. That is, as a function of the number of prototypes in $\mathcal{R}$ and number of retained eigenvectors in E. The feature curves are computed based on thirty repeated random 50%/50% data splits. In these splits, balanced class distributions are ensured by placing half the ROIs representing one class in the training set and the other half in the test set. In each split, the dimension range $N = [1, 2, \ldots, 30]$ is used in turn by selecting N random prototypes in the dissimilarity space approach and N positive eigenvectors in the embedding approach, in both cases from the training set.

Figure 4.6 shows the resulting prototype curves. QDC is more sensitive to the number of dimensions compared to LDC. This phenomenon can be explained by the increasing number of parameters in QDC, which requires more training samples for reliable estimation.



(a) Dissimilarity space using $D_{L_{HI}}$.

(b) Embedding of $D_{L_{HI}}$.

(c) Dissimilarity space using $D_{L_{EMD}}$.

(d) Embedding of $D_{L_{EMD}}$.

(e) Dissimilarity space using D_{L_2}.

(f) Embedding of D_{L_2}.

Figure 4.6: Feature curves of dissimilarity representation based LDC and QDC. Standard deviations are shown as dashed lines. The asterisks mark the minimum of each curve. The performance of the best kNN classifier, for $k = [1, \ldots, 5]$, using the training set as prototypes and the histogram dissimilarity in question as distance is also shown for reference as a horizontal line.

4.3.4 Classifier accuracy

The classification accuracy is evaluated using leave-one-out error estimation on the 168 ROIs, and the following classifier setups are evaluated:

- kNN using histogram dissimilarity measure L as distance. $k = [1, 2, \ldots, 5]$, $L = \{L_{HI}, L_{EMD}, L_2\}$.
- Classifier C in a dissimilarity space defined by random representation set selection from distance matrix D_L. $C = \{\text{LDC}, \text{QDC}\}$, $D_L = \{D_{L_{HI}}, D_{L_{EMD}}, D_{L_2}\}$.
- Classifier C in an embedding of a distance matrix D_L. $C = \{\text{LDC}, \text{QDC}\}$, $D_L = \{D_{L_{HI}}, D_{L_{EMD}}, D_{L_2}\}$.

The number of bins in the non-linear attenuation histogram is chosen as $N_b = \lfloor \sqrt[3]{N_p} \rfloor$, where N_p is the number of pixels in the ROI. In calculating L_{EMD}, the ground distance matrix, C in (4.2), is constructed such that the distance between two neighboring bins the attenuation histograms is one. More generally, the ground distance between bin i and bin j is $C_{ij} = |i - j|$. Further, we use the EMD implementation by Rubner [79]. The LDC and QDC class priors, $P(\omega_i)$ in (4.6) and (4.7), are estimated from data. The dimensionality of the dissimilarity spaces in all classifier setups is, somewhat arbitrarily, fixed to seven. All dissimilarity representation based classifiers perform reasonably well at this dimensionality according to the feature curves in Figure 4.6. The experiments are carried out in Matlab using the PRTools toolbox [26].

In general, all the classifiers perform well, see Table 4.1, with classification accuracies in the range $88.3\% - 97.0\%$. Using the dissimilarity space approach with randomly chosen prototypes generally performs worse than using kNN with histogram dissimilarity as distance directly. However, the embedding approach shows very promising results, especially when L_{EMD} is used as histogram dissimilarity. The best estimated classification accuracy of 97.0% is achieved using LDC in the approximate embedding of $D_{L_{EMD}}$, and this is significantly better than the best kNN with histogram dissimilarity as distance according to a McNemar's test [20] ($p = 0.046$).

4.4 Discussion and conclusions

The best dissimilarity representation based classifier achieves a classification accuracy of 97.0%, and this is significantly better ($p = 0.046$) than the best kNN classifier with histogram dissimilarity as distance, which achieved an accuracy of 92.9%. Generally, the embedding based classifiers perform slightly better than both the kNN and the dissimilarity space classifiers. Further, dissimilarity space based QDC, using only seven prototypes, performed similar to kNN. These results suggest that building classifiers in a dissimilarity representation, especially by embedding, is beneficial in the demonstrated application. The improved accuracy can be due to several factors. Firstly, a density based classifier built in a dissimilarity representation is more global, making use of all available training data in the classification decision, as opposed to

Classifier		L_{HI}	L_{EMD}	L_2
kNN using L as distance	1NN	91.7	92.9	92.3
	2NN	91.1	91.1	91.7
	3NN	92.9	91.1	92.3
	4NN	92.3	90.5	91.7
	5NN	91.1	89.3	91.1
Dissimilarity space	LDC	88.6 ($\pm$1.0)	88.3 ($\pm$1.7)	87.6 ($\pm$1.3)
	QDC	93.1 ($\pm$1.1)	90.1 ($\pm$2.0)	93.3 ($\pm$1.2)
Embedding	LDC	91.1	97.0	86.3
	QDC	94.1	95.2	95.2

Table 4.1: Results of the leave-one-out evaluation. The reported performance of the dissimilarity space experiments is an average of ten repeated leave-one-out experiments where the representation set is drawn at random each time. The same random representation set is used for all tested configurations. The standard deviations of these experiments are shown in parenthesis.

a kNN classifier, which classifies only based on the k nearest prototypes. Second, in the embedding, the classes seem to be approximately normal distributed, see Figure 4.5, which fits well with normal density based classifiers like LCD and QDC.

Accuracies previously reported in the literature on lung parenchyma classification in CT including at least one type of emphysema, and using measures of feature histograms as features in a feature space, are generally lower and lie in the range $76\% - 93,5\%$ [12, 65, 87, 102, 111]. These results are not directly comparable due to differences in the data, the choice of classes, etc. Nevertheless, the high accuracies of our experiments indicate that using the full feature histogram is beneficial and that a dissimilarity representation on histogram dissimilarities is a good way of utilizing the full feature histogram information.

In this chapter, we evaluated the dissimilarity space approach by drawing random prototypes for simplicity. However, prototype selection could be used instead, as in [69], which could improve the performance of the representation set approach. Another possibility would be to draw the prototypes at random on class-level such that an equal amount of prototypes from each class are present in the representation set.

QDC, and to some degree also LDC, showed unstable behavior in high dimensional dissimilarity spaces and embeddings, as seen in the feature curves in Figure 4.6. This problem could be addressed by regularizing the estimated covariance matrices, allowing a larger number of dimensions to be used [43]. This could possibly improve the classification accuracy.

A natural next step would be to try dissimilarity representations based on other

feature histograms than the attenuation histogram. For example, feature histograms describing local structure like local binary patterns [92] or other types of features previously used in lung parenchyma classification [12, 65, 87, 102, 111]. Combining the attenuation histogram and feature histograms describing local structure in a dissimilarity representation might improve performance.

In conclusion, we explore the use of normal density based classifiers built in a dissimilarity representation for lung parenchyma classification. Two different dissimilarity representation approaches are considered; embedding by classical scaling and the dissimilarity space approach, and dissimilarity representations based on different histogram dissimilarity measures are tried out. Two classifiers, LDC and QDC, are evaluated in the dissimilarity representations, and the best dissimilarity representation based classifier performed significantly better than using histogram dissimilarity directly as distance in a kNN classifier. A histogram dissimilarity representation allows for utilizing full feature histograms in classification, and through this representation, normal density based classifiers can be trained on histogram dissimilarity data. Further, sophisticated histogram dissimilarity measures, like the earth movers distance, fit naturally into this framework.

Chapter 5

Dissimilarity-Based Multiple Instance Learning

This chapter is based on the manuscript "Dissimilarity-Based Multiple Instance Learning," by L. Sørensen, M. Loog, D. M. J. Tax, W.-J. Lee, M. de Bruijne, and R. P. W. Duin, published in Proc. *Structural, Syntactic, and Statistical Pattern Recognition*, Lecture Notes in Computer Science, vol. 6218, pp. 129–138, 2010.

Abstract In this chapter, we propose to solve multiple instance learning problems using a dissimilarity representation of the objects. Once the dissimilarity space has been constructed, the problem is turned into a standard supervised learning problem that can be solved with a general purpose supervised classifier. This approach is less restrictive than kernel-based approaches and therefore allows for the usage of a wider range of proximity measures. Two conceptually different types of dissimilarity measures are considered: one based on point set distance measures and one based on the earth movers distance between distributions of within- and between set point distances, thereby taking relations within and between sets into account. Experiments on five publicly available data sets show competitive performance in terms of classification accuracy compared to previously published results.

5.1 Introduction

In multiple instance learning (MIL), complex objects are represented by sets of "sub-objects" where only the sets have an associated label, not the sub-objects. Following MIL terminology, the sets are termed bags and the sub-objects are termed instances. This kind of problem might, e.g., arise in medical image classification where a subject is known to suffer from a certain disease, but it is not clear exactly which regions in the associated medical image that contain the corresponding pathology. In this case, local image patches are the instances, the whole image is the bag, and the label of the bag is either ill or healthy.

The traditional approach to solving MIL problems involves explicit learning of a decision boundary in instance space that separates the instances capturing the concept from the remaining instances [21, 56]. A bag is then classified based on whether it contains an instance falling in this area. An alternative instance space approach involves labeling all instances with the same label as the bag they belong to. The problem is then treated as a standard supervised learning problem where all instances are classified in instance space, ultimately disregarding the multiple instance aspect of the original problem, and a bag is classified by combining the individual instance classifications in that bag [11].

The above mentioned approaches treat instances in the same bag independently in the learning step thereby disregarding potentially useful information. In some MIL problems, instances from the same bag collectively constitute that bag and should as such all contribute to the classification of that bag. Several authors have looked into using this information by applying learning at bag level with kernel-based methods. To name a few: Andrews *et al.* reformulated a support vector machine (SVM) optimization problem to operate directly on MIL problems at bag level [1]. Gärtner *et al.*, Tao *et al.*, and Zhou *et al.* designed specialized kernels for MIL problems and used standard SVMs with these kernels [33, 96, 114]. Chen *et al.* represented bags in an n-dimensional space where each dimension was the similarity between one of the n instances in the training set and the closest instance in a bag. Then a 1-norm SVM was used to simultaneously select the relevant features, or instances, and train a bag classifier [15].

In this chapter, we propose to use the dissimilarity representation approach to learning [68] for solving MIL problems at the bag level. Once the bag dissimilarity space has been constructed, the problem is turned into a standard supervised learning problem that can be solved with a general purpose supervised classifier. This is a proximity-based approach as are kernel-based methods, however, the dissimilarity representation approach does not require Mercer kernels as do kernel-based methods. A broader range of proximity measures, such as well known measures in pattern recognition like the Hausdorff distance and the single linkage distance, can therefore be used for solving MIL problems with this approach. We further propose, not only to consider all instances collectively in bag classification, but also to consider the relations among the instances within and between bags. This is similar in spirit to [114] where graphs capturing instance relations were constructed and used in a

SVM with a graph kernel [33]. A novel non-Mercer bag dissimilarity measure that is based on the earth movers distance (EMD) between instance distance distributions is proposed for this purpose. Compared to the graph kernel approach used in [114], the proposed bag dissimilarity measure is less rigid since distributions of instance distances are considered instead.

Dissimilarity-based learning has previously been applied in MIL. Wang and Zucker applied the k nearest neighbor (kNN) classifier to MIL problems by using the Hausdorff distance between the instances in two bags as the distance between these bags [107]. They showed that this was not sufficient to get good performance on the classical MIL data sets MUSK1 and MUSK2 [21], due to noise in the presence of negative instances in the positive bags, and suggested two adaptations of kNN instead. A key observation is that kNN using Hausdorff distance between instances is working on dissimilarities between bags, and one way of arriving at a more global and robust decision rule when dissimilarities between objects are available is via a dissimilarity representation [68]. Building a global classifier like the Fisher linear discriminant classifier (Fisher) on such a representation leads to a global decision rule that uses a weighted combination of the dissimilarities to all training set objects in classification. This means better utilization of the available training data, with possibly increased accuracy and less sensitivity to noise.

The rest of the chapter is organized as follows: Sections 5.2 and 5.3 briefly describe the MIL problem and the dissimilarity representation approach to learning. Section 5.4 presents two conceptually different types of dissimilarity measures between bags of instances. The first type is points set distance measures and the second type is based on EMD between distributions of instance distances within- and between bags. The proposed approach is evaluated by training and testing traditional supervised classifiers on dissimilarity representations of five publicly available MIL data sets. This is reported in Section 5.5. Finally, Section 5.6 provides a discussion and conclusions.

5.2 Multiple instance learning in short

In MIL [21], an object x_i is represented by a set, or bag, $B_i = \{\mathbf{x}_{ij}\}_{n_i}$ of n_i instances $\mathbf{x}_{ij}$, and a label $Y_i = \{+1, -1\}$ is associated with the entire bag. There are no labels y_{ij} associated directly with the instances, only indirectly via the label of the bag. This is different from standard supervised learning where objects are represented by a single instance, i.e., $B_i = \mathbf{x}_i$ and all instances therefore are directly labeled. The bag labels are interpreted in the following way in the original MIL formulation [21]: if $Y_i = -1$, then $\forall \mathbf{x}_{ij} \in B_i : y_{ij} = -1$. If $Y_i = +1$, then $\exists \mathbf{x}_{ij} \in B_i : y_{ij} = +1$. In other words, if a bag is labeled as positive, then at least one instance in that bag is a positive example of the underlying concept. This formulation can be relaxed to cope with a large and noisy set of instances by requiring that a positive bag contains a number or fraction of positive instances instead. In this chapter, we only consider two-class problems, but MIL can also be generalized to multi-class problems.

5.3 Dissimilarity representations in short

Objects x are traditionally represented by feature vectors in a feature vector space, and classifiers are built in this space. Alternatively, one can represent the objects by their pair-wise dissimilarities $d(x_i, x_j)$ and build classifiers on the obtained dissimilarity representation [68]. From the matrix of pair-wise object dissimilarities $D = [d(x_i, x_j)]_{n \times n}$ computed from a set of objects $\{x_1, \ldots, x_n\}$, there are different ways of arriving at a feature vector space where traditional vector space methods can be applied. In this chapter, we consider the dissimilarity space approach [68].

Given a training set $T = \{x_1, \ldots, x_n\}$, a subset $R = \{p_1, \ldots, p_k\} \subseteq T$ called the representation set containing prototype objects p_i is selected. An object x is represented with respect to R by the vector $D(x, R) = [d(x, p_1), \ldots, d(x, p_k)]$ of dissimilarities computed between x and the prototypes in R. This k-dimensional vector space based on R is called a dissimilarity space, and it is in this space that we propose to solve MIL problems at the bag level. In this chapter, we apply learning in the full dissimilarity space, i.e., $R = T$.

5.4 Bag dissimilarity space

The idea we propose is to map the bags into a dissimilarity space $D(\cdot, R = \{B_i\}_k)$. Here the bags are represented as single objects, positioned with respect to their dissimilarities to the prototype bags in R. In this space, the MIL problem can be considered as a standard supervised classification problem where each object $x_i = B_i$ has label Y_i and general purpose supervised classifiers can be directly applied. The separation of the bags in the obtained dissimilarity space is very much dependent on the choice of bag dissimilarity measure $d(B_i, B_j)$. In the following, we present two conceptually different types of dissimilarity measures for bags of instances.

5.4.1 Point set distance measures

The instances $\mathbf{x}$ reside in a common space and bags B can therefore be thought of as sets of objects in this space. In the case of vectorial instances, these objects are points in a vector space. This leads to the idea of computing dissimilarities between bags using point set distance measures. In this chapter, we experiment with the minimum distance

$$d_{min}(B_i, B_j) = \min_{p,q} ||\mathbf{x}_{ip} - \mathbf{x}_{jq}||_2 \tag{5.1}$$

and the Hausdorff distance

$$d_H(B_i, B_j) = \max\{d_{dir}(B_i, B_j), d_{dir}(B_j, B_i)\} \tag{5.2}$$

which is based on the directed distance $d_{dir}(B_i, B_j) = \max_p \min_q ||\mathbf{x}_{ip} - \mathbf{x}_{jq}||_2$. These point set distance measures were also used in a modified kNN classifier in [107].

Both point set distance measures (5.1) and (5.2) use the distance between two single instances in the end. These measures may therefore be sensitive to noisy instances, and they are in general insensitive to the number of positive instances in a

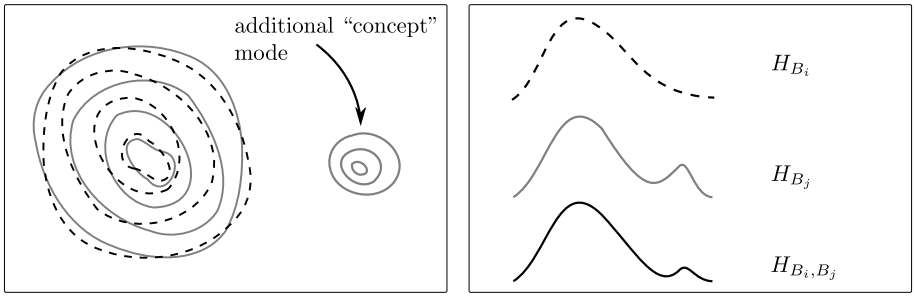


Figure 5.1: Left: Illustration of two similar bag class distributions where one of the distributions, typically the positive bag distribution, has an extra mode corresponding to the positive instances. **Right**: Distributions of instance distances, from top to bottom: within bag instance distances in a bag from the class with no additional mode, typically the negative class; within bag instance distances in a bag from the class with an additional mode, typically the positive class. Notice the extra "bump" in the distribution; instance distances between two bags, one from each class.

positive bag. This may not be desirable when constructing a bag dissimilarity representation, and taking more information about the instances in a bag into account in the bag dissimilarity measure may lead to a better representation of the bags.

5.4.2 Measures based on between- and within bag instance distances

Zhou *et al.* conjectured that instances in a bag are rarely independently and identically distributed and that relations among the instances may convey important information when applying learning at bag level [114]. In a similar spirit, we propose two bag dissimilarity measures that take relations among instances into account, or more precisely, the distribution of instance distances within a bag and between bags. It is assumed that the instances in the two bag classes follow distributions in the common instance space that are very similar, with the slight difference that one distribution contains additional modes capturing the concept(s). This situation is illustrated, for one additional mode, to the left in Figure 5.1. This could, e.g., be the situation in a MIL problem in medical image classification where the positive medical images contain lesions surrounded by healthy tissue whereas the negative images only contain healthy tissue. The additional mode in one of the bag class distributions gives rise to an extra "bump" in the distribution of instance distances within bags from that class, compared to bags from the other class, as illustrated to the right in Figure 5.1. Further, the bump can also be seen in the histogram of instance distances computed between bags from the two classes.

We propose to use the within bag instance distance histograms H_{B_i} and H_{B_j}, computed from bag B_i and B_j, respectively, and the between bag instance distance

histogram H_{B_i,B_j}, computed between bag B_i and B_j. The bag dissimilarity measure is then computed as the pair-wise histogram dissimilarity $d_{i,ij} = d(H_{B_i}, H_{B_i,B_j})$. $d_{i,ij}$ can be seen as the directed dissimilarity from B_i to B_j. The maximum and the mean of the directed dissimilarities from each of the two bags are proposed as two symmetric dissimilarity measures for bags

$$d_{BWmax}(B_i, B_j) = \max\{d_{i,ij}, d_{j,ij}\} \tag{5.3}$$

and

$$d_{BWmean}(B_i, B_j) = \frac{1}{2}(d_{i,ij} + d_{j,ij}). \tag{5.4}$$

The histogram dissimilarities are computed using EMD [80] between the normalized empirical distributions. For one-dimensional histograms $H = [h_1, \ldots, h_n]^T$ and $K = [k_1, \ldots, k_n]^T$ of equal number of bins n and equal mass, EMD can be computed as the L1-norm between the cumulative histograms of H and K: $d_{EMD}(H, K) = \sum_{i=1}^{n} |\sum_{j \leq i} h_j - \sum_{j \leq i} k_j|$.

5.4.3 A second dissimilarity space

Initial experiments showed that linear classifiers performed poorly when built on the obtained bag dissimilarity representations whereas the nearest neighbor classifier (1NN) performed quite well. This indicates that the bags are separated in the obtained dissimilarity representations, but that the decision boundaries between the positive bags and the negative bags are complicated and non-linear, and/or that the class distributions are multi-modal in these new representations. An extra preprocessing step is therefore done before applying linear classifiers. From $D(\cdot, X)$ computed on the full data set X, a new dissimilarity representation $D2$ is constructed such that $D2(x_i, x_j) = ||D(x_i, X) - D(x_j, X)||_2, \forall_{x_i,x_j} \in X$. The linear classifiers are built on this representation. This is a transductive learning approach since all objects are used to construct the representation D2. It is, however, important to note that the labels of the objects are not considered in this construction. Tao *et al.* also used transductive learning to solve MIL problems [96].

5.5 Experiments and results

The proposed approach is evaluated on the two standard data sets in MIL, namely MUSK1 and MUSK2 originally used in [21], and on three recently published image retrieval data sets [1].

5.5.1 MUSK1 and MUSK2

These are the standard MIL data sets, and they consist of descriptions of aromatic molecules that have been labeled according to whether they smell "musky" or not. A bag represents a molecule, and the instances in a bag are low energy shapes of the molecule described by 166-dimensional feature vectors. The MUSK1 data set

comprises 47 positive bags and 45 negative bags, and each bag is represented by 2 to 40 instances. The MUSK2 data set comprises 39 positive bags and 63 negative bags, and each bag is represented by 1 to 1044 instances. The data was obtained from the UCI Machine Learning Repository [5], and we refer to this source as well as to [21] for further information about the data.

5.5.2 Image retrieval

This data comprises three data sets that are subsets of the Corel data set. Each data set consists of 100 positive bags, or example images; elephant, fox, or tiger, and 100 negative bags, or background images, which are randomly drawn from a pool of photos of other animals. Each image is represented by 2-13 instances (apart from a single image in the tiger data set that is represented by a single instance), which are 230-dimensional feature vectors describing the color, texture and shape in subsegments of the image. The data was obtained from the homepage[1] associated with [1] and we refer to these sources for further information about the data.

5.5.3 Evaluation

The proposed dissimilarity representations are evaluated by training and testing three supervised classifiers on the bags in the given dissimilarity space. These classifiers are: 1NN; SVM with a linear kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ where $\mathbf{x}_i = D2(B_i, X)$ and trade-off parameter $C = 1$; and Fisher. For 1NN and Fisher we use the pattern recognition toolbox PRTools [26], and for SVM we use LIBSVM [14].

Classification accuracies are estimated using leave-one-out and 10-fold cross-validation, since these are commonly used performance measures in the MIL literature [1, 11, 21, 33, 56, 107, 114]. 10-fold cross-validation is sometimes performed once and sometimes the average of a repeated number of 10-fold cross-validation procedures is reported. Here we perform one 10-fold cross-validation. The results are presented in Table 5.1 and Table 5.2 where also previously published results are reported.

The classification accuracies of 1NN are quite close to the ones previously reported in the literature. The high 1NN classification accuracies on the MUSK1 and MUSK2 data set indicate that the bags are well separated in the obtained bag dissimilarity space defined by D. Fisher performs poorly when built on D with an average classification accuracy of 62.1% whereas SVM performs decent when built on D with an average classification accuracy of 78.4%. However, building them on a second dissimilarity representation $D2$ constructed from D, as described in Section 5.4.3, improves performance considerably for Fisher with an average absolute increase of 19.3% and slightly for SVM with an average absolute increase of 4%. 1NN performs slightly worse when applied to $D2$ compared to D, and the numbers reported in Table 5.1 and Table 5.2 for 1NN are therefore based on D. SVM and Fisher generally perform better than 1NN. We also tried kNN with k optimized using cross-validation on the training set in each fold which achieved similar performance to 1NN.

[1] http://www.cs.columbia.edu/~andrews/mil/datasets.html

Classifier	Bag dissimilarity measure		MUSK1	MUSK2
1NN (on D)	d_{min}	(5.1)	90.2 / **91.3**	86.9 / 84.6
	d_H	(5.2)	88.0 / 87.9	86.1 / 82.5
	d_{BWmax}	(5.3)	85.8 / 86.9	82.8 / 77.7
	d_{BWmean}	(5.4)	89.1 / 91.2	85.3 / 80.7
SVM (on $D2$)	d_{min}	(5.1)	90.0 / 90.1	92.2 / 87.5
	d_H	(5.2)	88.0 / 88.0	91.2 / 85.5
	d_{BWmax}	(5.3)	89.1 / 89.0	82.2 / 88.3
	d_{BWmean}	(5.4)	**91.2** / 89.0	85.3 / 85.0
Fisher (on $D2$)	d_{min}	(5.1)	90.1 / 90.1	**93.5** / **92.7**
	d_H	(5.2)	88.0 / 86.9	90.3 / 88.2
	d_{BWmax}	(5.3)	90.1 / 87.9	87.7 / 87.4
	d_{BWmean}	(5.4)	**91.2** / 91.2	89.8 / 90.3
citation-kNN [107]			**92.4** / -	86.3 / -
iterated discrim APR [21]			- / **92.4**	- / 89.2
diverse density [56]			- / 88.9	- / 82.5
mi-SVM [1]			- / 87.4	- / 83.6
MI-SVM [1]			- / 77.9	- / 84.3
SVM polynomial minimax kernel [33]			**92.4** / -	86.3 / -
SVM MI kernel [33]			87.0 / -	92.2 / -
MILES [15]			86.3 / 87.0	87.7 / **93.1**
$k_\wedge$ *emph* transduction [96]			- / 91.2	- / 90.3
$k_{\wedge/\vee}$ *emph* transduction [96]			- / 90.2	- / 92.2
MIGraph [114]			- / 90.0	- / 90.0
miGraph [114]			- / 88.9	- / 90.3

Table 5.1: Classification accuracy on the MUSK1 and MUSK2 data set, reported as leave-one-out / ten-fold cross-validation. Accuracies reported in the literature are shown in the bottom part of the table. Cases in the literature where the classification accuracy is not reported using leave-one-out or ten-fold cross-validation are marked with "-". The highest accuracy among the dissimilarity representation-based classifiers as well as the highest accuracy in general is marked in boldface in each column.

Classifier	Bag dissimilarity measure		elephant	fox	tiger
1NN (on D)	d_{min}	(5.1)	78.0 / 78.0	60.0 / 59.5	77.0 / 74.0
	d_H	(5.2)	70.0 / 69.5	52.0 / 50.0	67.0 / 64.5
	d_{BWmax}	(5.3)	75.0 / 77.5	57.5 / 57.0	68.0 / 66.0
	d_{BWmean}	(5.4)	80.0 / 79.0	59.5 / 59.0	70.5 / 71.5
SVM (on $D2$)	d_{min}	(5.1)	85.5 / 83.5	**67.5** / 65.0	77.5 / 78.0
	d_H	(5.2)	84.0 / 84.5	37.5 / 49.0	73.5 / 73.5
	d_{BWmax}	(5.3)	**89.0** / **89.0**	64.5 / 56.0	69.5 / 62.0
	d_{BWmean}	(5.4)	87.0 / 87.0	62.5 / 58.5	78.0 / 76.5
Fisher (on $D2$)	d_{min}	(5.1)	86.0 / 84.5	66.0 / **66.0**	78.5 / 78.0
	d_H	(5.2)	84.5 / 85.0	59.0 / 59.0	73.5 / 72.0
	d_{BWmax}	(5.3)	88.5 / 88.5	66.5 / 63.0	81.0 / 78.5
	d_{BWmean}	(5.4)	**89.0** / 88.5	64.5 / 64.0	**81.5** / **79.5**
mi-SVM [1]			- / 82.2	- / 58.2	- / 78.9
MI-SVM [1]			- / 81.4	- / 59.4	- / 84.0
MIGraph [114]			- / 85.1	- / 61.2	- / 81.9
miGraph [114]			- / 86.8	- / 61.6	- / **86.0**

Table 5.2: Classification accuracy on the image retrieval data. See the caption of Table 5.1 for further details.

Across all five data sets, SVM and Fisher built on dissimilarity representations show excellent performance. On the MUSK1 and MUSK2 data set, the classifiers achieve accuracies close to the best reported accuracies in the literature. On the image retrieval data sets, SVM with a linear kernel, as well as Fisher, perform better than the SVM's adapted to MIL problems [1] in two out of three data sets. This indicates that taking instance relations into account is beneficial in this kind of problems, as is also seen in [114].

5.6 Discussions and conclusions

The linear classifiers built on the proposed dissimilarity representations performed better than the best results in the MIL literature in some cases, and in the remaining cases close to the best published results [1,15,21,33,56,96,107,114]. It should be noted that the classifiers were applied "off the shelf" and that, e.g., the trade-off parameter C in SVM was not tuned by cross-validation but fixed to 1. Also, the classifiers were trained and tested in dissimilarity spaces of dimension equal to the number of training samples. This is no problem for SVM. For Fisher, the pseudo-inverse was used. It may be possible to obtain even better results than the ones reported in Table 5.1 and Table 5.2 by proper regularization or by reducing the dimensionality of the dissimilarity space, e.g., by prototype selection [69].

SVM shows worse than random performance on some of the image retrieval data sets, in particular when built on the dissimilarity representation obtained using the Hausdorff distance, d_H, on the fox data set. This could be caused by a strong class overlap in the dissimilarity space. This is also indicated by the fact that both 1NN and Fisher perform worse on this representation compared to the other representations.

The minimum point set distance, d_{min}, works well as bag dissimilarity measure. Similar results were reported in [107]. This is somewhat surprising since classes are expected to be overlapping in MIL due to positive bags also containing negative instances. The explanation is that the distribution of the positive instances is more dense compared to the negative instances in the used data sets, and therefore a bag containing at least one positive instance is more likely to be close to another bag containing at least one positive instance than to a bag containing only negative instances.

To conclude, we have shown that the dissimilarity representation approach can be used to solve MIL problems. Global decision rules in the form of general purpose supervised linear classifiers built in a bag dissimilarity space achieves excellent classification accuracies on publicly available MIL data sets. The approach is general, and we see this as a promising direction that allows for using a wider range of proximity measures between bags in solving MIL problems compared to the popular kernel-based approaches. Further, there are indications that taking relations among instances into account improves the performance on certain MIL problems, such as the image retrieval problems.

Chapter 6

Dissimilarity-Based Multiple Instance Learning for COPD Quantification

This chapter is based on the manuscript "Image Dissimilarity-Based Quantification of Lung Disease from CT," by L. Sørensen, M. Loog, P. Lo, H. Ashraf, A. Dirksen, R. P. W. Duin, and M. de Bruijne, published in Proc. *Medical Image Computing and Computer Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 6361, pp. 37–44, 2010.

Abstract In this chapter, we propose to classify medical images using dissimilarities computed between a collection of regions of interest. The images are mapped into a dissimilarity space using an image dissimilarity measure, and a standard vector based classifier is applied in this space. The classification output of this approach can be used in computer aided-diagnosis problems where the goal is to detect the presence of abnormal regions or to quantify the extent and severity of abnormalities in these regions. The proposed approach is applied to quantify chronic obstructive pulmonary disease in computed tomography (CT) images, achieving an area under the receiver operating characteristic curve of 0.817. This is significantly better compared to combining individual region classifications into an overall image classification, and compared to common computerized quantitative measures in pulmonary CT.

6.1 Introduction

Quantification of abnormality in medical images often involves classification of regions of interest (ROIs), and combination of individual ROI classification outputs into one global measure of disease for the entire image [3,58,65,78,89,93,105]. These measures may, e.g., express a probability of the presence of certain abnormalities or reflect the amount or severity of disease.

A global image measure based on the fusion of several independent ROI classifications disregards the fact that the ROIs belong to a certain image in the classification step. Moreover, in some cases only global image labels are available, while the images are still represented by ROIs in order to capture localized abnormalities. In some studies, this is handled by propagating the image label to the ROIs within that image, which again allows fusion of individual ROI classifications, to obtain a global image measure [3, 78, 89]. However, an image showing abnormality will generally comprise both healthy and abnormal regions, and the above approach, incorrectly, labels ROIs without abnormality in such an image as abnormal.

In this chapter, we propose to classify medical images using dissimilarities computed directly between the images, where the images are represented by a collection of regions. In this approach, all ROIs in an image contribute when that image is compared to other images, thereby taking into account that the ROIs collectively constitute that image. Further, problems where only a global image label is available are handled automatically since the classification is done at the image level. The images are mapped into a dissimilarity space [68] in which a standard vector space-based classifier can be directly applied, and the soft output of this classifier is used as quantitative measure of disease. The measure used to compute the dissimilarity between two images is the crucial component in this approach, and we evaluate five different image dissimilarity measures in the experiments.

The proposed approach is applied to quantify chronic obstructive pulmonary disease (COPD) in volumetric pulmonary computed tomography (CT) images using texture. Several general purpose classifiers built in the obtained image dissimilarity spaces are evaluated and compared to image classification by fusion of individual ROI classifications as was used in [89].

6.2 Image dissimilarity space

We propose to represent a set of images $\{I_1, \ldots, I_n\}$ by their pair-wise dissimilarities $d(I_i, I_j)$ and build classifiers on the obtained dissimilarity representation [68]. From the matrix of pair-wise image dissimilarities $D = [d(I_i, I_j)]_{n \times n}$ computed from the set of images, there exist different ways of arriving at a feature vector space where traditional vector space methods can be applied. In this chapter, we consider the dissimilarity space approach [68]. An image dissimilarity space is constructed of dimension equal to the size of the training set $|T| = |\{J_1, \ldots, J_m\}| = m$, where each dimension corresponds to the dissimilarity to a certain training set image J. All images I are represented as single points in this space, and are positioned according

to their dissimilarities to the training set images $D(I,T) = [d(I,J_1),\ldots,d(I,J_m)]$. The image dissimilarity measure is a function from two images, represented as sets of ROIs, to a non-negative scalar $d(\cdot,\cdot) : \mathcal{P}(S) \times \mathcal{P}(S) \to \mathbb{R}_+$ where S is the set of ROIs and $\mathcal{P}(S)$ is the power set of S. It is in this part of the proposed approach that the ROIs are taken collectively into account.

6.3 Image dissimilarity measures

The main issue in obtaining the image dissimilarity space, is the definition of $d(\cdot,\cdot)$. Since the application in this chapter is quantification of COPD in pulmonary CT images based on textural appearance in the ROIs, we will focus on image dissimilarity measures suitable for this purpose. In texture-based classification of lung tissue, the texture is sometimes assumed stationary [65, 78, 89, 93]. We will make the same assumption and, therefore, disregard the spatial location of the ROIs within the lungs. The following are then desirable properties of an image dissimilarity measure for quantification of abnormality:

1. Spatial location within the image does not matter. ROIs should be compared solely based on the textural appearance within those regions.
2. The amount of diseased tissue does matter. An image with many abnormal regions is more diseased than an image with few abnormal regions.
3. The appearance of abnormal tissue does matter. Two images with abnormal regions of the same size but with different types of abnormality should be considered different.

A simple and straightforward image dissimilarity measure between two images, I_1 and I_2, having the above properties is the sum of all pair-wise ROI dissimilarities:

$$d_{sum}(I_1,I_2) = \sum_{i,j} \Delta(\mathbf{x}_{1i},\mathbf{x}_{2j}) \tag{6.1}$$

where $\mathbf{x}_{1i}$ is the i'th ROI in I_1 and $\Delta(\cdot,\cdot)$ is a texture appearance dissimilarity measure between two ROIs. However, when all ROIs in one image are compared to all ROIs in the other image, the discriminative information of abnormality present in only a few ROIs may be lost. One way to avoid this is to match every ROI in one image with the most similar ROI in the other image. This is the minimum sum distance [29]:

$$d_{ms}(I_1,I_2) = \sum_i \min_j \Delta(\mathbf{x}_{1i},\mathbf{x}_{2j}) + \sum_j \min_i \Delta(\mathbf{x}_{2j},\mathbf{x}_{1i}). \tag{6.2}$$

However, this image dissimilarity measure allows several ROIs in one image to be matched with the same ROI in the other image. This may not be desirable for quantifying abnormality since an image with a small abnormal area is considered similar to an image with a large abnormal area. The image dissimilarity measure

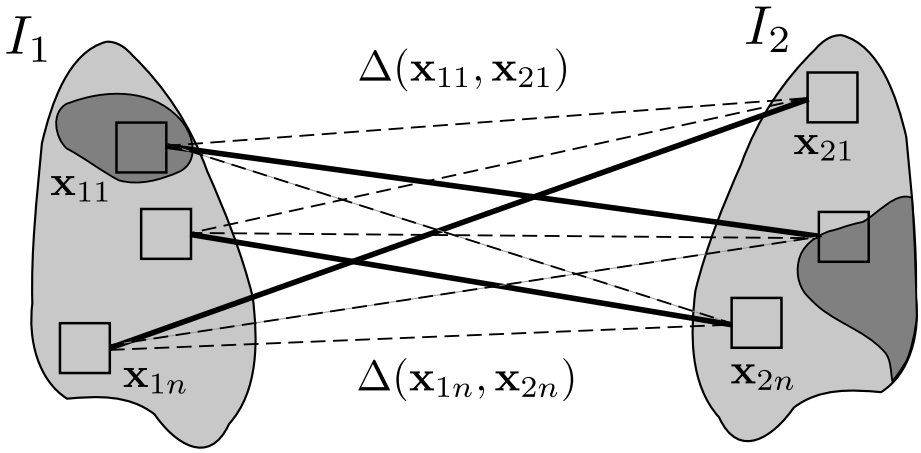


Figure 6.1: Illustration of the graph considered when computing the dissimilarity between two images, I_1 and I_2, in (6.4). All edges have an associated weight $\Delta(\mathbf{x}_{1i}, \mathbf{x}_{2j})$ that expresses the textural dissimilarity between the two corresponding ROIs $\mathbf{x}_{1i}$ and $\mathbf{x}_{2j}$. The edges in the perfect matching with minimum weight M^* are shown as solid lines, and the remaining edges, not in M^*, are shown as dashed lines.

proposed in the following is a trade-off between d_{sum} and d_{ms}; it is the sum of several pair-wise ROI dissimilarities, where only one-to-one matchings are allowed, thereby considering images with a small abnormal area as dissimilar to images with a large abnormal area.

6.3.1 Bipartite graph matching-based image dissimilarity measure

The dissimilarity between two images, or sets of ROIs, $I_1 = \{\mathbf{x}_{1i}\}_n$ and $I_2 = \{\mathbf{x}_{2i}\}_n$, can be expressed as the minimum linear sum assignment between the two sets according to $\Delta(\cdot, \cdot)$. This can be seen as assigning the ROIs in one set to the ROIs in the other set in a way such that the two sets are as similar as possible while only allowing one-to-one matchings. Let $G = (I_1 \cup I_2, E)$ be a weighted undirected bipartite graph with node sets I_1 and I_2 where $|I_1| = |I_2| = n$, edge set $E = \{\{\mathbf{x}_{1i}, \mathbf{x}_{2j}\} : i, j = 1, \ldots, n\}$, and with weight $\Delta(\mathbf{x}_{1i}, \mathbf{x}_{2j})$ associated with each edge $\{\mathbf{x}_{1i}, \mathbf{x}_{2j}\} \in E$. The resulting graph is illustrated in Figure 6.1. A subset M of E is called a perfect matching, or assignment, if every node of G is incident with exactly one edge from M. The perfect matching with minimum weight M^* is given by

$$M^* = \underset{M}{\operatorname{argmin}} \sum_{(\mathbf{x}_{1i}, \mathbf{x}_{2j}) \in M} \Delta(\mathbf{x}_{1i}, \mathbf{x}_{2j}) : M \text{ is a perfect matching.} \tag{6.3}$$

This problem can be solved efficiently using the Hungarian algorithm [50]. The resulting image dissimilarity measure is thus

$$d_{la}(I_1, I_2) = \sum_{(\mathbf{x}_{1i}, \mathbf{x}_{2j}) \in M^*} \Delta(\mathbf{x}_{1i}, \mathbf{x}_{2j}) \tag{6.4}$$

where M^* is obtained via (6.3). No normalization is needed since the images contain an equal amount of ROIs, i.e., n ROIs. Although not used in this chapter, the formulation can also be relaxed to handle images containing a varying number of ROIs. This will result in an image dissimilarity measure that does not obey the triangle inequality due to partial matches of images. However, this is no problem in the dissimilarity space appraoch.

6.4 Experiments

6.4.1 Data

The data consists of 296 low-dose volumetric CT images from the Danish Lung Cancer Screening Trial with the following scan parameters: tube voltage 120 kV, exposure 40 mAs, slice thickness 1 mm, and in-plane resolution ranging from 0.72 to 0.78 mm. 144 images are from subjects diagnosed as healthy and 152 images are from subjects diagnosed with moderate to very severe COPD. Both groups are diagnosed according to spirometry [77].

6.4.2 Evaluation

The image dissimilarity-based approach is applied by building classifiers in the CT image dissimilarity spaces using $d(\cdot, \cdot)$. This is compared to using $d(\cdot, \cdot)$ directly as distance in a k nearest neighbor classifier (kNN), which for $k = 1$ corresponds to template matching, and to fusing individual ROI classifications for image classification [89]. A posterior probability for each image being positive is obtained using leave-one-out estimation, and receiver operating characteristic (ROC) analysis is used to evaluate the different methods by means of the area under the ROC curve (AUC). The CT image dissimilarity spaces considered in each leave-out trial are of dimension equal to the size of the training set, i.e., 295-dimensional.

Apart from the three image dissimilarity measures described in Section 6.3, (6.1), (6.2), and (6.4), we also experiment with the Hausdorff distance [29], d_h. This is a classical point set distance measure that do not obey the second property described in Section 6.3, since it ultimately rely on the dissimilarity between two single ROIs, or points, one from each image. Thus, a total of four different CT image dissimilarity representations are considered in the experiments, one based on each of the four image dissimilarity measures d_{sum}, d_{ms}, d_{la}, and d_h.

6.4.3 Classifiers

All CT images are represented by a set of 50 ROIs, of size $41 \times 41 \times 41$ voxels, that each are described by three filter response histograms capturing the local image texture. The filters are: Laplacian of Gaussian (LG) at scale 0.6 mm, gradient magnitude (GM) at scale 4.8 mm, and Gaussian curvature (GC) at scale 4.8 mm. The ROI size as well as the filters are selected based on the results in [89]. The ROI dissimilarity measure used in all experiments is based on the L1-norm between filter response histograms: $\Delta(\mathbf{x}_1, \mathbf{x}_2) = L_1(h_{LG}(\mathbf{x}_1), h_{LG}(\mathbf{x}_2)) + L_1(h_{GM}(\mathbf{x}_1), h_{GM}(\mathbf{x}_2)) + L_1(h_{GC}(\mathbf{x}_1), h_{GC}(\mathbf{x}_2)))$ where $h_i(\mathbf{x})$ is the response histogram of filter i computed in ROI $\mathbf{x}$.

A SVM with a linear kernel and trade-off parameter $C = 1$ is applied in the obtained CT image dissimilarity spaces. kNN is applied in the following three ways: in the image dissimilarity spaces, using the image dissimilarities directly as distance, and using ROI dissimilarity directly for ROI classification followed by fusion. $k = 1$ is used as well as $k = \sqrt{n}$ where n is the number of prototypes [46]. When classifying CT images, this is $k = \lfloor\sqrt{295}\rfloor = 17$, and when classifying ROIs, this is $k = \lfloor\sqrt{(295 \times 50)}\rfloor = 121$. The following combination rules are considered for fusing individual ROI classifications into image classifications: quantile-based fusion schemes with quantiles ranging from 0.01, i.e., the minimum rule, to 1.00, i.e., the maximum rule, and the mean rule [55]. We also compare to two common densitometric measures in pulmonary CT, namely, relative area of emphysema (RA) and percentile density (PD) using the common thresholds of -950 Hounsfield units (HU) and 15% respectively [108]. These measures are computed from the entire lung fields and are denote RA_{950} and PD_{15}.

6.4.4 Results

Table 6.1 shows the estimated AUCs for all the classifiers. The best CT image-dissimilarity based classifier, SVM built in CT image dissimilarity space using d_{la}, achieves an AUC of 0.817. This is better than the best performing mean rule ROI fusion-based classifier, 121NN, which achieves an AUC of 0.751. The common densitometric measures, RA_{950} and PD_{15}, perform worse than all the texture based classifiers. The quantile-based fusion schemes only performed better than the mean rule in one case, 121NN using maximum rule achieved an AUC of 0.757, and they are therefore not reported in Table 6.1. SVM in image dissimilarity space using d_{la} or d_{sum} is significantly better, with $p = 0.0028$ and $p = 0.0270$, respectively, than 121NN using the mean rule, while SVM using d_{ms} is not, with $p = 0.085$, according to DeLong, DeLong, and Clarke-Pearson's test [17].

6.5 Discussion

Image dissimilarity measures that match each ROI of one image to an ROI of the other image, under some restrictions, are expected to work well for quantification

	in image dissimilarity space			using $d(\cdot,\cdot)$ directly	
	SVM	1NN	17NN	1NN	17NN
d_h	0.609	0.522	0.624	0.566	0.668
d_{sum}(6.1)	0.793	0.619	0.643	0.504	0.663
d_{ms} (6.2)	0.795	0.632	0.725	0.600	0.768
d_{la} (6.4)	**0.817**	0.612	0.671	0.593	0.741

fusion of ROI classifications	
1NN	0.721
121NN	**0.751**
RA_{950}	0.585
PD_{15}	0.589

Table 6.1: AUCs for COPD diagnosis. **Left:** The results of classification in image dissimilarity space, as well as using the image dissimilarities directly in kNN. **Right:** The results of fusion of individual ROI classification outputs for image classification using the mean rule. The best performing classifier in each approach is marked in bold-face.

of abnormality within the proposed framework, mainly because more information is taken into account, but also due to increased robustness to noisy ROIs. This is in contrast to measures relying on the match between two ROIs only, such as the Hausdorff distance that is included in the experiments for the sake of completeness. Further, the main arguments for building a more global decision rule, such as SVM, in a dissimilarity space instead of applying kNN using the dissimilarity directly as distance are: better utilization of the training data and therefore reduced sensitivity to noisy prototypes [68]. This may explain why SVM with a linear kernel built in the dissimilarity space obtained using d_{la} is the best performing of the CT image dissimilarity-based approaches. However, validation on an unseen data set would be needed to draw a final conclusion on this. The experiments showed that SVM with a linear kernel built in the CT image dissimilarity space obtained using d_{la} performed significantly better than using kNN for ROI classification together with the mean rule for CT image classification ($p < 0.05$). This implies that performing the classification at image level, taking into account that an image is in fact a collection of ROIs that collectively constitute that image, is beneficial compared to classifying ROIs individually, while disregarding the fact that they do belong to a certain image.

The computational complexity of the proposed approach using either of the image dissimilarities (6.1), (6.2), or (6.4), in terms of the number of times $\Delta(\cdot,\cdot)$ is evaluated in order to classify a CT image, is the same compared to using the image dissimilarities directly as distance in kNN and to fusion of ROI classifications that are classified using kNN. All approaches require a total of $50 \times 50 \times 295$ evaluations of $\Delta(\cdot,\cdot)$ for classification of a CT image.

When an image is represented by a collection of ROIs while only a global label for the entire image is available, the problem of classifying the image can be formulated as a multiple instance learning (MIL) problem [21]. Fusion of independent ROI classifications in order to arrive at an overall image classification can be seen as a "simple" algorithm for solving such a problem. In this chapter, we propose to use the dissimilarity-based approach of Pekalska *et al.* [68] on image dissimilarities for solving MIL problems in medical imaging. The approach is similar in spirit to various kernel-

based MIL algorithms, such as [33]. The dissimilarity-based approach, however, puts less restrictions on the proximity measure used for comparing objects. Kernel-based approaches require the kernel to be positive definite, which excludes well-known proximity measures such as the Hausdorff distance [29] as well as the image dissimilarity measure proposed in this chapter. Within our framework such measures can be used without any problem.

In conclusion, dissimilarities computed directly between medial images, where the images are represented by a collection of ROIs, was proposed for image classification. This is an alternative to fusion of individual ROI classifications within the images. A SVM built in a dissimilarity space using an image dissimilarity measure based on a minimum sum perfect matching in a weighted bipartite graph, with ROIs as nodes and the textural dissimilarity between two ROIs as edge weight, achieved an AUC of 0.817 on a COPD quantification problem in volumetric pulmonary CT.

Chapter 7

Summary and General Discussion

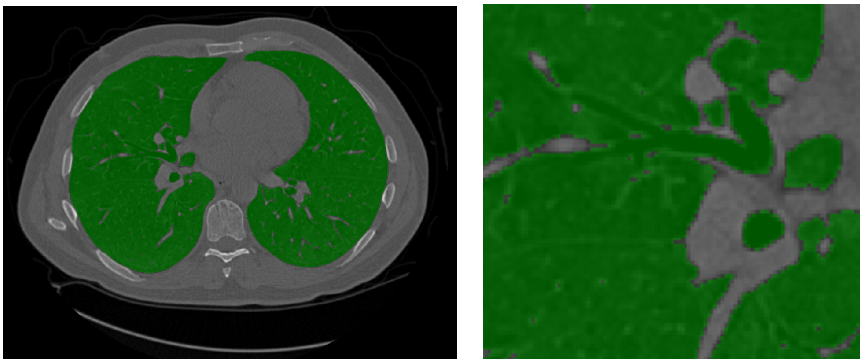
This thesis describes several learning-based methods for quantitative analysis of emphysema and/or chronic obstructive pulmonary disease (COPD) in computed tomography (CT) images of the lungs. These methods may also be applicable to other lung diseases and modalities.

7.1 Summary

In **Chapter 2**, a method for quantitative analysis of emphysema using textural information in CT images of the lungs is presented. Various sources of variability in CT, including difference in inspiration level, scanner drift, and noise artifacts, as well as the subtle and diffuse appearance of emphysema, makes this a challenging task. We propose to tackle this task using supervised learning with texture descriptors computed in manually annotated regions of interest (ROIs) containing examples of the patterns of interest, resulting in data-driven a rule for quantitative analysis that is able to cope with noise and varying patterns. The used texture descriptor is the joint histogram of intensity and rotation-invariant local binary patterns (LBPs). Two emphysema classes are defined according to existing emphysema subtype definitions together with a healthy tissue class. The two emphysema classes are: centrilobular emphysema (CLE), defined as multiple small low-attenuation areas; and paraseptal emphysema (PSE), defined as multiple low-attenuation areas in a single layer along the pleura often surrounded by interlobular septa that is visible as thin white walls. The method is trained on a set of 168 manually annotated ROIs and is subsequently applied to classify all lung parenchyma pixels in three CT slices from 39 different subjects. The pixel classification probabilities for the healthy class are used as a reciprocal measurement of the amount of emphysema present, and the probabilities are, for each subject, fused into an overall quantitative measure that correlates significantly better with a pulmonary function test (PFT) than does relative area of emphysema (RA).

LBP cannot be straightforwardly applied in three dimensions, although several approximations have been proposed [112], and since volumetric CT is becoming more

and more common, this is a potential limitation. However, the whole learning framework can be applied in volumetric CT by using another texture descriptor. **Chapter 3** presents a fully automatic, data-driven approach for training a supervised classifier for quantitative analysis of COPD. The classification scheme is similar to that of Chapter 2, however, with the difference that we use histograms of filter responses from a multi-scale, rotation invariant Gaussian filter bank as texture descriptor, and apply the learning framework to volumetric CT images. The main objective of the chapter is to illustrate that it is possible to use a learning-based approach for quantitative analysis without any human intervention, thereby obtaining an objective measure that is not limited to current knowledge and experience of experts. Instead of manually annotating ROIs in an image, ROIs are sampled at random from within an automatic segmentation of the lung fields and labeled according to PFTs, the current gold standard for COPD diagnosis [77], associated with the sampled CT image. The segmented lung fields contain both the lung parenchyma and the lumen as well as part of the wall of the airways, see Figure 7.1, and both main components of COPD [77, 83], namely, emphysema and chronic bronchitis, are therefore targeted. The method is evaluated on a set of 296 volumetric CT images and is shown to be significantly better at COPD diagnosis than RA and percentile density (PD). The resulting quantitative measure of COPD correlates significantly with a PFT whereas RA and PD does not. It is further shown that the proposed method is less sensitive to the inspiration level of a subject during CT scanning – a major source of variability in CT – compared to RA and PD.



(a) Axial slice overlayed with a lung fields segmentation.

(b) Zoom-in on the middle right part of the axial slice in (a).

Figure 7.1: Example of a lung fields segmentation used in this thesis. (a) Axial slice from a CT image overlayed with a lung fields segmentation (shown in transparent green). (b) Zoom in on the axial slice overlayed with a lung fields segmentation in (a). Notice that both the airway lumen and part of the airway walls, as well as the fissure, are in the segmentation.

Chapter 4 investigates dissimilarity representation-based classifiers for classification of healthy and emphysematous lung parenchyma. The classification schemes

in Chapters 2 and 3 already use a dissimilarity-based classifier for ROI classification, a k nearest neighbor (kNN) classifier with dissimilarities between ROIs directly as distances. However, kNN is a local decision rule that does not use the full training set in classification. This information can be incorporated in classification by representing the ROIs in a dissimilarity representation and by training a normal density-based classifier on the training data in this representation. ROIs are classified in this representation with the trained classifier, resulting in a more global decision rule that uses the full training set in classification. Two dissimilarity representation approaches are considered, dissimilarity space and embedding using classical scaling where negative eigenvalues and corresponding eigenvectors are disregarded [70]. In the dissimilarity space approach, the dissimilarity measure is interpreted as a mapping to a feature space where each dimension corresponds to the dissimilarity to a certain prototype object. This approach can be straightforwardly applied since ROI dissimilarities are used directly. If a good ROI dissimilarity measure is derived, which is equivalent to defining good features in a traditional feature space approach, a space defined by characteristic objects of each class is expected to separate the data well. In the second approach, the data is embedded in a feature space by finding a, possibly lower dimensional, configuration of the data for which distances approximate the original dissimilarities between objects. Both a linear and a quadratic discriminant classifier is applied using each representation. The method is evaluated for ROI classification using a fairly simple texture descriptor, namely, the intensity histogram, on the same data set with 168 ROIs as is used in Chapter 2. It is shown that the classification accuracy of the best performing dissimilarity representation-based classifier is significantly better compared to kNN using ROI dissimilarities directly as distance.

In Chapter 3, PFTs were used to label CT images. These labels were subsequently propagated to the ROIs from each image in order to apply supervised learning on the ROIs. A CT image was classified by combining ROI classification outcomes in that image. This approach can be seen as a simple and implicit way of solving the so-called multiple instance learning (MIL) problem. That is, a problem where an object, the CT image, is represented by a set of sub-objects, the ROIs, with only a global label available, the PFTs. The aforementioned approach disregards set membership in sub-object classification. MIL is a well studied problem in pattern recognition and machine learning where several algorithms taking explicitly into account that sub-objects comes from a certain set have been proposed [113]. These algorithms can roughly be divided into algorithms that either operate on sub-object level or on set level. **Chapter 5** describes a novel MIL algorithm that operates on set level. The general idea is to use the dissimilarity space approach, also used in Chapter 4 for ROI classification, on the sets of sub-objects by defining a suitable dissimilarity measure between two sets that takes the sub-objects into account. The method is evaluated on five publicly available MIL benchmark data sets. The performance of the proposed MIL algorithm, in terms of classification accuracy, is competitive with previously published results in the literature. We also investigate the benefit of taking relations between the sub-objects into account by using a novel set dissimilarity measures based on the earth movers distance between distributions of within and between set sub-

object distances. This improves performance on some of the benchmark data sets, and this approach may in general be beneficial when certain patterns co-exist in some sets in the data.

The MIL algorithm presented in Chapter 5 operates on set level, and this may also be applied for classification of CT images of the lungs where the images are represented by ROIs and only a global CT image label is available. In this way, it is explicitly taken into account that an ROI comes from a certain CT image which may contain both healthy and diseased regions. The essential component of the MIL algorithm presented in Chapter 5 is the measure used for computing dissimilarity between sets. In **Chapter 6**, we propose to classify CT images directly using a suitable dissimilarity measure computed between CT images where the CT images are represented by a set of ROIs. In this way, the classification is performed directly at CT image level, instead of combination of ROI classification results for classification of CT images as done in Chapters 2 and 3. Several CT image dissimilarity measures for quantitative analysis of COPD are investigated, including a novel linear assignment-based CT image dissimilarity measure. The basic idea is, that we, in a global sense, want to match two images in the best way possible according to the textural patterns in the ROIs. This can be achieved by formulation the problem as a linear assignment problem, a well known problem in the areas of combinatorial optimization and operations research that can be solved in polynomial time in the number of ROIs using special purpose algorithms such as the Hungarian algorithm [50]. The ROIs of two CT images being compared are viewed as nodes in a weighted bipartite graph where the edge weights are the textural dissimilarities between the connected ROIs. The perfect matching with minimum weight in this graph, defined as the summed edge weights over a set of edges that connects each node of the graph with exactly one edge, is the proposed dissimilarity that expresses the optimal matching of the two images. We evaluate the CT image dissimilarity measures, both by using the dissimilarity directly as distance in a kNN classifier and by using the MIL algorithm of Chapter 5 on the same data set with 296 volumetric CT images as is used in Chapter 3. A support vector machine (SVM) built on the data in the obtained CT image dissimilarity space is significantly better at COPD diagnosis, both compared to combining individual ROI classifications into an overall image classification, as is done in Chapter 3, and compared to RA and PD.

7.2 Computerized quantitative measures

A learning-based approach using texture features is a good choice for quantitative analysis of COPD in CT, when sufficient labeled training data is available to train a classifier. The labeled training data can either be obtained by manual annotation, which is the common approach, or by using meta-data, such as PFTs. The investigated learning-based approaches outperform the common computerized quantitative measures in the clinical literature, RA and PD, as demonstrated in Chapters 2, 3, and 6. We expect the same to hold for other computerized measures in the clinical literature that rely only on a parameter derived from the histogram of voxel intensities,

as well. Examples of other measures than the ones compared to in this thesis are: mean lung density [38, 83], total lung volume [38], lung weight [38], the mode of the histogram [83]. Common characteristics for all these measures are: 1) they rely solely on intensity, or density, information; 2) they use individual voxel information; 3) only voxel resolution is considered; and 4) they summarize the global voxel intensity histogram by a single parameter. By using texture descriptors, such as histograms of Gaussian derivative-based filter responses, to characterize an ROI, or a voxel based on the surrounding ROI, much more of the available information is taken into account. In particular: 1 and 2) local structure is taken into account, which also implies that relations among voxels are considered; 3) multi-scale approaches consider the information at scales larger than voxel resolution; and 4) the lung is summarized by several local texture measurements, one for each ROI considered, that in turn is local information summarized by full histograms. The advantages of using more information from the CT images, in the form of texture, is first of all better discriminative ability leading to a better quantitative measure. But also less sensitivity to the inspiration level of the subject during the CT scanning as demonstrated in Chapters 2 and 3. However, all this comes at a price. A learning-based method is needed in order to turn the computed texture features into a quantitative measure for COPD, e.g., supervised learning, as is done in Chapters 2 and 3, or MIL, as is done in Chapter 6. First of all, there is the aforementioned need of labeled training data. However, one also ends up with a much more complicated decision rule that is not easily interpretable, e.g., compared to simple rules such as RA where the number of voxels with attenuation close to that of air relative to the total amount of voxels in the lung is computed [83]. Questions such as "what do we actually measure in the image?" are therefore not straightforward to answer for a learning-based approach using texture features. The learned decision rule may also deteriorate in performance when applied on new data with different characteristics than the training data. This situation may, e.g., arise when a learning-based approach is trained on a data set from one scanner and applied to a data set from the same scanner but with different scan parameters, such as radiation dose and reconstruction kernel, or from a different scanner.

Overall, two conceptually different learning-based approaches are presented for texture-based quantitative analysis of COPD in CT images in this thesis. The first approach, described in Chapters 2 and 3, classifies ROIs within an image and the classification results are combined into an overall image classification by posterior fusion. The second approach, described in Chapter 6, classifies the images directly using an image dissimilarity measure that is based on the ROIs within the images. Both the first approach, when applied as in Chapter 3, where global CT image labels are propagated to the associated ROIs, and the second approach can be seen as MIL algorithms. For which situations one approach is preferable to the other remains an open question that should be further investigated. Although, the results of Chapter 6 indicate that classifying CT images directly provides a more reliable indicator of pulmonary function. It should be noted, however, that the DLCST population, where the CT images are from, is "relatively healthy". The subjects had to be able to climb 2 flights of stairs (36 steps) without pausing and have a corrected forced expiratory

volume in 1 second (FEV_1%pred) of at least 30% in order to be allowed entry into the study [67], and since we use baseline CT images from DLCST in Chapter 6, this indication only holds for subjects that are not too diseased. The first approach is preferable when the location of abnormalities within a CT image is needed, e.g., in the form of a probability map of disease. It is not directly possible to extract this information from the second approach.

The computational complexity, in terms of the number ROI dissimilarity computations needed in order to classify a test CT image, is the same for the two approaches. Assuming that the same ROI dissimilarity measure is used in the two approaches, that each CT image is represented by m ROIs, and that there are n CT images in the training set. In the first approach, an ROI is classified by computing the dissimilarity to all prototype ROIs in the kNN classifier, requiring mn computations, and this is done for all ROIs in the test image resulting in a total of m^2n ROI dissimilarity computations. In the second approach, the dissimilarity between the test CT image and a training CT image requires m^2 ROI dissimilarity computations. The dissimilarity needs to be computed between the test image and all training images in order to represent the test image in the CT image dissimilarity space, resulting in a total of m^2n ROI dissimilarity computations.

7.3 Applications

The proposed learning-based approaches in this thesis are directly applicable in situations where a large data set is available and an objective quantitative measure for specifically that data is needed, as long as labeled training data can be obtained. This could be in pharmaceutical studies or in clinical research, for example. The approaches can be trained on a sub-set of the data and applied to the remaining data, e.g., in a cross-validation procedure. However, the learned decision rule is tied to data with the same or similar characteristics. Alternatively, techniques for dealing with different data such as transfer learning techniques [64] could be considered in cases where the learned decision rule is applied to new data with substantially different characteristics, but this is outside the scope of this thesis. Application in clinical practice is less straightforward due to large variety in data, mainly caused by different scanning protocols. Implementation in a work station aimed at clinical use would at least require a reliable segmentation of the lung fields and an agreed upon set of base filters. Based on the results of the conducted experiments in Chapter 2 and 3 where sequential forward feature selection (SFS) is used to select a small sub-set of filters from a large pool of possible filters, it could be the following base filters: gradient magnitude, Laplacian of the Gaussian, and preferably also the Gaussian function itself. The specific scale at which the different base filters are applied can be tuned according to the scanner settings, either automatically, e.g., using SFS, or by setting them according to scan parameters such as the used reconstruction kernel and radiation dose. Small scale features may not work well in low dose CT due to noise, for example. The lung fields segmentation can be obtained using algorithms already implemented in commercial work stations or alternatively using the automatic

segmentation algorithm used in Chapter 3.

7.4 Improvements

The texture descriptors used in this thesis are all rotation invariant. The main motivation behind this choice is that we are dealing with a complex task of classifying diffuse and heterogenous pathology, and taking out rotation limits the number of possible patterns making the learning task easier. This is of course provided that rotation-invariance does not discard important discriminative information. Looking at the achieved performance of the methods, see Chapters 1 and 2, for example, rotation-invariant descriptors does work well in this setting. It may still be the case, however, that discriminative information is lost. For example, CLE is predominantly in the upper lobes and the formation of this pattern as well as the shape of emerging bullae may to some extent be aligned with the orientation of the nearby airway tree. Capturing this information with a rotation variant texture descriptor may therefore improve the discriminative capability of the complete system. Further, spatial information is not considered in this work, but pathology, such as CLE in the upper lobes, may be detectable as structure with different orientation than the general orientation of the parenchyma in a particular region. Combining rotation variant texture descriptors with spatial location according to an anatomical coordinate system may therefore improve the discriminative capability. The anatomical coordinate system could, e.g., be according to the overall orientation of the lungs or according to the local orientation of the airway tree. We suspect that rotation variant texture descriptors, possibly in conjunction with spatial location, may be useful. However, this would probably require more training data, due to the larger possible variations in patterns and increase of dimensionality, in order to reach the same performance as when using rotation invariant texture descriptors.

Both main components of COPD, i.e., emphysema and chronic bronchitis, are implicitly targeted in the quantitative measures for COPD proposed in Chapters 3 and 6. This is because the segmentation of the lung fields, used to indicate which voxels to use when computing texture descriptors, includes both the lung parenchyma, the small airways, and part of the lumen as well as part of the wall of the larger airways, see Figure 7.1. However, emphysema is still expected to contribute most to the COPD measure. A recent study shows that chronic bronchitis and emphysema make independent contributions to the airflow limitation in COPD [66], and it would therefore be interesting to combine the proposed measures with more explicit measures of airway disease. For example, with measures that are based on a segmentation of the airway tree. This could be common measures such as the internal area or the wall area [8], or more recent measures such as the airway wall thickness at a lumen perimeter of 10 mm (Pi10) [66] or the normalized wall intensity sum [71]. In general, one could also imagine combining the proposed learning-based measures with other markers for COPD, apart from measures from airway tree segmentations, for an improved marker. For example, meta-data not used to train the learning-based measures. This could be done by adding the meta-data as extra features, improved

performance has previously been demonstrated by adding age as feature [18], or by classifier combination of the proposed measure with the output of a separate classifier that uses the meta-data as features, for example.

Chapter 4 demonstrated that it is possible to increase ROI classification performance using a dissimilarity representation approach as compared to using ROI dissimilarities directly as distance in a kNN classifier, possibly due to better utilization of the training data. We therefore suspect that using this approach for ROI classification in Chapters 2 and 3, instead of a kNN classifier, will also increase the CT image classification performance obtained by fusion of ROI posteriors and provide a better quantitative measure.

7.5 Future prospects

MIL problems often arise naturally in medical image classification. For example, when human experts label complete images by judging the images as a whole, yet the images comprise several sub-regions that are all taken into account in this judgement. Classification of images containing at least one lung nodule versus images containing no nodules is one example of a problem that can be formulated and solved as a MIL problem. There are, nevertheless, relatively few published medical image analysis studies using MIL [9, 28, 32, 49, 90, 110]. It would therefore be interesting to further explore the possible benefit of using MIL in medical imaging, including quantitative analysis of COPD in CT, which to our knowledge has only been done in [89, 90], the publications that Chapters 3 and 6 are based on, respectively. The MIL algorithm that is described in Chapter 5 and applied to the lung data in Chapter 6 is novel. The method is less restrictive on the proximity measure used between sets, as compared to the kernel-based MIL methods in the literature, and it is therefore possible to use a broader range of dissimilarity measures. The method may therefore prove advantageous over existing MIL methods in some cases, and the proposed algorithm should be further evaluated on other available benchmark MIL data sets, with different properties, e.g., different instance distributions, in order to investigate this. Further, development of new set dissimilarity measures, e.g., measures that take relations between instances within a set into account, as the one proposed in Chapter 5, may improve the performance of MIL solutions for problems where the instance distribution has a certain structure, e.g., a different instance distribution in bags of different classes.

To this date, there exist no publicly available CT image data sets with labeled examples of emphysema and/or COPD, or other lung diseases, that can be used for evaluation and comparison of different methods for computerized quantitative analysis of lung disease. Published studies rely on their own private data sets. Also, no extensive surveys exist that compare different published methods on the same, preferably large, data set. It is therefore not possible to make a final conclusion on which texture features and which classifier to use. Based on the results of this thesis, and on the results of Chapter 2 in particular, we recommend using histograms of filter responses from a multi-scale, rotation invariant Gaussian filter bank (GFB)

or histograms of LBPs jointly with intensity, over moments of GFB or measures on run-length matrices and co-occurrence matrices, which are often used in the literature [12,65,75,102,103,111]. However, it still remains an open question which features and which classifier is most suitable for quantitative analysis of emphysema and/or COPD in CT. A more elaborate comparison of different methods, or even better, establishment of publicly available data sets with labeled data that can serve as a reference standard could aid in answering these questions and advance the field. This could, e.g., be via public challenges similar to recent challenges in other fields of medical image analysis carried out at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). Examples of recent challenges are: the Extraction of Airways from CT challenge (EXACT09) and the Volume Change Analysis of Nodules challenge (VOLCANO09), that both were part of The Second International Workshop on Pulmonary Image Analysis[1] during MICCAI in 2009.

The study using supervised learning presented in Chapter 3, which was applied to 296 volumetric CT images from 296 different subjects, as well as the application of the developed MIL algorithm on the same data, presented in Chapter 6, are to our knowledge the largest learning-based studies using texture features in CT for classification of lung abnormality. Previous studies in CT are applied to $34 - 116$ subjects, either represented by volumetric CT images or by one or more CT slices each [12, 44, 87, 100, 102, 111]. The mentioned studies use manually annotated data for training a supervised classifier. The proposed methods in Chapters 3 and 6 use PFTs for acquiring labels and can therefore be trained on large data sets without much effort, as long as there is also PFT data available. Alternatively, other meta-data could be used, e.g., other markers or risk factors for COPD such as biomarkers measured in blood samples [84] or smoking history. These could be used in isolation to define new classification problems, or in conjunction with PFTs in order to arrive at a more confident labeling of the CT images. Using other meta-data than PFTs has not been explored in this thesis. The volumetric CT image data set used in Chapters 3 and 6 is a sub-set of the Danish Lung Cancer Screening Trial (DLCST) database [67]. DLCST is a longitudinal screening trial comprising more than 9500 volumetric CT images with associated PFTs from former or current smokers that have been scanned for five consecutive years. In the future, we plan to apply the developed methodology to the entire database. This has very exciting aspects. First of all, the data size of the proposed future application is more than an order of magnitude larger than any previously published learning-based studies using texture features in CT for classification of lung abnormality. Application to the full DLCST database will result in a much better estimate of the performance, the reproducibility, etc., of the learning-based methods of Chapters 3 and 6, which facilitates better comparison, both between the two approaches and related to existing densitometric measures such as RA and PD. Secondly, training on such a large data set is expected to result in better parameter estimates in the methods that better captures the true underlying distribution of the classes in the data set, and as a consequence a better

[1] http://www.lungworkshop.org/2009/

objective quantitative measure of COPD. It may even be possible to reliably learn parameters of more complex approaches than the ones proposed in the thesis. Thirdly, the interpretation and evaluation of the results of applying the proposed approaches to the full DLCST database may lead to an increased understanding of the disease mechanisms of COPD, e.g., a more precise picture of how the disease is spatially distributed within the lungs and how it evolves as a function of time. Knowledge about the distribution of pathology is, e.g., relevant for patients considered for lung-volume-reduction surgery, and it has been shown that the anatomical distribution of emphysema has a major impact on the outcome of such a procedure [30]. The decreased sensitivity to inspiration level of the proposed measures, compared to RA and PD, also facilitates more reliable identification of risk factors for COPD.

Bibliography

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, pages 561–568. MIT Press, 2002.

[2] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[3] Y. Arzhaeva, L. Hogeweg, P. A. de Jong, M. A. Viergever, and B. van Ginneken. Global and local multi-valued dissimilarity-based classification: application to computer-aided detection of tuberculosis. In G.-Z. Yang, D. J. Hawkes, D. Rueckert, J. A. Noble, and C. J. Taylor, editors, *Medical Image Computing and Computer Assisted Intervention*, volume 5762 of *Lecture Notes in Computer Science*, pages 724–731. Springer, 2009.

[4] H. Ashraf, P. Lo, S. B. Shaker, M. de Bruijne, A. Dirksen, P. Tønnesen, M. Dahlbäck, and J. H. Pedersen. Change in smoking habits affects lung density by CT. In *American Thoracic Society International Conference*, 2009.

[5] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

[6] A. A. Bankier, V. De Maertelaer, C. Keyzer, and P. A. Gevenois. Pulmonary emphysema: subjective visual grading versus objective quantification with macroscopic morphometry and thin-section CT densitometry. *Radiology*, 211(3):851–858, 1999.

[7] P. J. Barnes. Chronic obstructive pulmonary disease. *The New England Journal of Medicine*, 343(4):269–280, 2000.

[8] P. Berger, V. Perot, P. Desbarats, J. M. Tunon de Lara, R. Marthan, and F. Laurent. Airway wall thickness in cigarette smokers: quantitative thin-section CT assessment. *Radiology*, 235(3):1055–1064, 2005.

[9] J. Bi and J. Liang. Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE Computer Society Press, 2007.

[10] E. Bruno, N. Moënne-Loccoz, and S. Marchand-Maillet. Asymmetric learning and dissimilarity spaces for content-based retrieval. In H. Sundaram, M. R. Naphade, J. R. Smith, and Y. Rui, editors, *International Conference on Image and Video Retrieval*, volume 4071 of *Lecture Notes in Computer Science*, pages 330–339. Springer, 2006.

[11] A. Cannon and D. Hush. Multiple instance learning using simple classifiers. In *International Conference on Machine Learning and Applications*, pages 123–128. IEEE Computer Society Press, 2004.

[12] F. Chabat, G.-Z. Yang, and D. M. Hansell. Obstructive lung diseases: texture classification for differentiation at CT. *Radiology*, 228(3):871–877, 2003.

[13] H.-P. Chan, L. Hadjiiski, C. Zhou, and B. Sahiner. Computer-aided diagnosis of lung cancer and pulmonary embolism in computed tomography–a review. *Academic Radiology*, 15(5):535 – 555, 2008.

[14] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[15] Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

[16] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[17] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[18] A. Depeursinge, J. Iavindrasana, G. Cohen, A. Platon, P.-A. Poletti, and H. Muller. Lung tissue classification in HRCT Data integrating the clinical context. In *Annual IEEE Symposium on Computer-Based Medical Systems*, pages 542–547. IEEE Computer Society Press, 2008.

[19] A. Depeursinge, D. Sage, A. Hidki, A. Platon, P.-A. Poletti, M. Unser, and H. Muller. Lung tissue classification using wavelet frames. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6259–6262, 2007.

[20] T. G. Dietterich. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[21] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[22] A. Dirksen, N. H. Holstein-Rathlou, F. Madsen, L. T. Skovgaard, C. S. Ulrik, T. Heckscher, and A. Kok-Jensen. Long-range correlations of serial FEV1 measurements in emphysematous patients and normal subjects. *Journal of Applied Physiology*, 85(1):259–265, 1998.

[23] K. Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4-5):198–211, 2007.

[24] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[25] R. P. W. Duin, D. de Ridder, and D. M. J. Tax. Featureless pattern classification. *Kybernetika*, 34(4):399–404, 1998.

[26] R. P. W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, and D. M. J. Tax. PRTools, a Matlab toolbox for pattern recognition. http://www.prtools.org, 2004.

[27] R. P. W. Duin and D. M. J. Tax. Classifier conditional posterior probabilities. In A. Amin, D. Dori, P. Pudil, and H. Freeman, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, volume 1451 of *Lecture Notes in Computer Science*, pages 611–619. Springer, 1998.

[28] M. M. Dundar, G. Fung, B. Krishnapuram, and R. B. Rao. Multiple-instance learning algorithms for computer-aided detection. *IEEE Transactions on Biomedical Engineering*, 55(3):1015–1021, 2008.

[29] T. Eiter and H. Mannila. Distance measures for point sets and their computation. *Acta Informatica*, 34(2):109–133, 1997.

[30] A. Fishman, F. Martinez, K. Naunheim, S. Piantadosi, R. Wise, A. Ries, G. Weinmann, D. E. Wood, and National Emphysema Treatment Trial Research Group. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *The New England Journal of Medicine*, 348(21):2059–2073, 2003.

[31] O. Friman, M. Borga, M. Lundberg, U. Tylén, and H. Knutsson. Recognizing emphysema - a neural network approach. In *International Conference on Pattern Recognition*, pages 512–515. IEEE Computer Society Press, 2002.

[32] G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao. Multiple instance learning for computer aided diagnosis. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, pages 425–432. MIT Press, 2006.

[33] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In C. Sammut and A. G. Hoffmann, editors, *International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 2002.

[34] M. L. Giger, H.-P. Chan, and J. Boone. Anniversary paper: History and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM. *Medical Physics*, 35(12):5799–5820, 2008.

[35] L. Goldfarb. A unified approach to pattern recognition. *Pattern Recognition*, 17(5):575–582, 1984.

[36] G. A. Gould, W. MacNee, A. McLean, P. M. Warren, A. Redpath, J. J. Best, D. Lamb, and D. C. Flenley. CT measurements of lung density in life can quantitate distal airspace enlargement–an essential defining feature of human emphysema. *American Review of Respiratory Disease*, 137(2):380–392, 1988.

[37] G. A. Gould, A. T. Redpath, M. Ryan, P. M. Warren, J. J. Best, D. C. Flenley, and W. MacNee. Lung CT density correlates with measurements of airflow limitation and the diffusing capacity. *European Respiratory Journal*, 4(2):141–146, 1991.

[38] H. Guenard, M. H. Diallo, F. Laurent, and J. Vergeret. Lung density and lung mass in emphysema. *Chest*, 102(1):198–203, 1992.

[39] J. W. Gurney. Pathophysiology of obstructive airways disease. *Radiologic Clinics of North America*, 36(1):15–27, 1998.

[40] J. W. Gurney, K. K. Jones, R. A. Robbins, G. L. Gossman, K. J. Nelson, D. Daughton, J. R. Spurzem, and S. I. Rennard. Regional distribution of emphysema: correlation of high-resolution CT with pulmonary function tests in unselected smokers. *Radiology*, 183(2):457–463, 1992.

[41] R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[42] S. Hu, E. A. Hoffman, and J. M. Reinhardt. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE Transactions on Medical Imaging*, 20(6):490–498, 2001.

[43] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[44] N. Kim, J. B. Seo, Y. Lee, J. G. Lee, S. S. Kim, and S.-H. Kang. Development of an automatic classification system for differentiation of obstructive lung disease using HRCT. *Journal of Digital Imaging*, 22(2):136–148, 2009.

[45] M. Kinsella, N. L. Müller, R. T. Abboud, N. J. Morrison, and A. DyBuncio. Quantitation of emphysema by computed tomography using a "density mask" program and correlation with pulmonary function tests. *Chest*, 97(2):315–321, 1990.

[46] J. Kittler and F. M. Alkoot. Moderating k-NN classifiers. *Pattern Analysis and Applications*, 5(3):326–332, 2002.

[47] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[48] H. Knutsson and C.-F. Westin. Normalized and differential convolution: Methods for interpolation and filtering of incomplete and uncertain data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 515–523. IEEE Computer Society Press, 1993.

[49] B. Krishnapuram, J. Stoeckel, V. C. Raykar, R. B. Rao, P. Bamberger, E. Ratner, N. Merlet, I. Stainvas, M. Abramov, and A. Manevitch. Multiple-instance learning improves CAD detection of masses in digital mammography. In E. A. Krupinski, editor, *International Workshop on Digital Mammography*, volume 5116 of *Lecture Notes in Computer Science*, pages 350–357. Springer, 2008.

[50] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

[51] E. Levina and P. Bickel. The earth mover's distance is the mallows distance: some insights from statistics. In *IEEE International Conference on Computer Vision*, volume 2, pages 251–256. IEEE Computer Society Press, 2001.

[52] A. Løkke, P. Lange, H. Scharling, P. Fabricius, and J. Vestbo. Developing COPD: a 25 year follow up study of the general population. *Thorax*, 61(11):935–939, 2006.

[53] P. Lo, J. Sporring, H. Ashraf, J. J. H. Pedersen, and M. de Bruijne. Vessel-guided airway tree segmentation: A voxel classification approach. *Medical Image Analysis*, 14(4):527–538, Aug 2010.

[54] P. Lo, J. Sporring, J. J. H. Pedersen, and M. de Bruijne. Airway tree extraction with locally optimal paths. In G.-Z. Yang, D. J. Hawkes, D. Rueckert, J. A. Noble, and C. J. Taylor, editors, *Medical Image Computing and Computer Assisted Intervention*, volume 5762 of *Lecture Notes in Computer Science*, pages 51–58. Springer, 2009.

[55] M. Loog and B. van Ginneken. Static posterior probability fusion for signal detection: applications in the detection of interstitial diseases in chest radiographs. In *International Conference on Pattern Recognition*, pages 644–647. IEEE Computer Society Press, 2004.

[56] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*. The MIT Press, 1997.

[57] A. H. Mir, M. Hanmandlu, and S. N. Tandon. Texture analysis of CT images. *IEEE Engineering in Medicine and Biology Magazine*, 14(6):781–786, 1995.

[58] N. L. Müller, C. A. Staples, R. R. Miller, and R. T. Abboud. "Density mask". An objective method to quantitate emphysema using computed tomography. *Chest*, 94(4):782–787, 1988.

[59] K. Murphy, B. van Ginneken, E. M. van Rikxoort, B. J. de Hoop, M. Prokop, P. Lo, M. de Bruijne, and J. P. W. Pluim. Obstructive pulmonary function: patient classification using 3D registration of inspiration and expiration CT images. In M. Brown, M. de Bruijne, B. van Ginneken, A. Kiraly, J.-M. Kuhnigk, C. Lorenz, K. Mori, and J. Reinhardt, editors, *Proc. of The Second International Workshop on Pulmonary Image Analysis*, pages 37–47, 2009.

[60] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[61] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[62] A. Oliver, X. Lladó, J. Freixenet, and J. Martí. False positive reduction in mammographic mass detection using local binary patterns. In N. Ayache, S. Ourselin, and A. J. Maeder, editors, *Medical Image Computing and Computer Assisted Intervention*, volume 4791 of *Lecture Notes in Computer Science*, pages 286–293. Springer, 2007.

[63] P. Paclik and R. P. W. Duin. Dissimilarity-based classification of spectra: computational issues. *Real-Time Imaging*, 9(4):237–244, 2003.

[64] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints), 2009.

[65] Y. S. Park, J. B. Seo, N. Kim, E. J. Chae, Y. M. Oh, S. D. Lee, Y. Lee, and S.-H. Kang. Texture-based quantification of pulmonary emphysema on high-resolution computed tomography: Comparison with density-based quantification and correlation with pulmonary function test. *Investigative Radiology*, 43(6):395–402, 2008.

[66] B. D. Patel, H. O. Coxson, S. G. Pillai, A. G. N. Agustí, P. M. A. Calverley, C. F. Donner, B. J. Make, N. L. Müller, S. I. Rennard, J. Vestbo, E. F. M. Wouters, M. P. Hiorns, Y. Nakano, P. G. Camp, P. V. N. Fauerbach, N. J. Screaton, E. J. Campbell, W. H. Anderson, P. D. Paré, R. D. Levy, S. L. Lake, E. K. Silverman, D. A. Lomas, and International COPD Genetics Network. Airway wall thickening and emphysema show independent familial aggregation in chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, 178(5):500–505, 2008.

[67] J. J. H Pedersen, H. Ashraf, A. Dirksen, K. Bach, H. Hansen, P. Toennesen, H. Thorsen, J. Brodersen, B. G. Skov, M. Døssing, J. Mortensen, K. Richter, P. Clementsen, and N. Seersholm. The danish randomized lung cancer CT screening trial–overall design and results of the prevalence round. *Journal of Thoracic Oncology*, 4(5):608–614, 2009.

[68] E. Pekalska and R. P. W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition Letters*, 23(8):943–956, 2002.

[69] E. Pekalska, R. P. W. Duin, and P. Paclík. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2):189–208, 2006.

[70] E. Pekalska, P. Paclík, and R. P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, 2001.

[71] J. Petersen, P. Lo, M. Nielsen, G. Edula, H. Ashraf, A. Dirksen, and M. de Bruijne. Quantitative analysis of airway abnormalities in CT. In N. Karssemeijer and R. M. Summers, editors, *Medical Imaging: Computer-Aided Diagnosis. Proceedings of SPIE.*, volume 7624, page 6241S. SPIE Press, 2010.

[72] M. Petrou and P. G. Sevilla. *Image Processing: Dealing With Texture.* Wiley, 2006.

[73] T. L. Petty. COPD in perspective. *Chest*, 121:116S–120S, 2002.

[74] M. Prasad, A. Sowmya, and I. Koch. Designing relevant features for continuous data sets using ICA. *International Journal of Computational Intelligence and Applications*, 7(4):447–468, 2008.

[75] M. Prasad, A. Sowmya, and P. Wilson. Multi-level classification of emphysema in HRCT lung images. *Pattern Analysis and Applications*, 12(1):9–20, 2009.

[76] P. H. Quanjer, G. J. Tammeling, J. E. Cotes, O. F. Pedersen, R. Peslin, and J. C. Yernault. Lung volumes and forced ventilatory flows. Report working party standardization of lung function tests, European Community for Steel and Coal. Official statement of the European Respiratory Society. *The European Respiratory Journal. Supplement*, 16:5Ű–40, 1993.

[77] K. F. Rabe, S. Hurd, A. Anzueto, P. J. Barnes, S. A. Buist, P. Calverley, Y. Fukuchi, C. Jenkins, R. Rodriguez-Roisin, C. van Weel, J. Zielinski, and Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American Journal of Respiratory and Critical Care Medicine*, 176(6):532–555, 2007.

[78] J. Raundahl, M. Loog, P. Pettersen, L. B. Tanko, and M. Nielsen. Automated effect-specific mammographic pattern measures. *IEEE Transactions on Medical Imaging*, 27(8):1054–1060, 2008.

[79] Y. Rubner. Code for the earth movers distance (EMD). http://ai.stanford.edu/ rubner/emd/default.htm, 1998.

[80] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[81] M. Schwaiblmair, T. Beinert, M. Seemann, J. Behr, M. Reiser, and C. Vogelmeier. Relations between cardiopulmonary exercise testing and quantitative high-resolution computed tomography associated in patients with alpha-1-antitrypsin deficiency. *European Journal of Medical Research*, 3(11):527–532, 1998.

[82] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, Dec 1979.

[83] S. B. Shaker, A. Dirksen, K. S. Bach, and J. Mortensen. Imaging in chronic obstructive pulmonary disease. *COPD*, 4(2):143–161, 2007.

[84] S. B. Shaker, K. A. von Wachenfeldt, S. Larsson, I. Mile, S. Persdotter, M. Dahlbäck, P. Broberg, B. Stoel, K. S. Bach, M. Hestad, T. E. Fehniger, and A. Dirksen. Identification of patients with chronic obstructive pulmonary disease (COPD) by measurement of plasma biomarkers. *The Clinical Respiratory Journal*, 2(1):17–25, 2008.

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

[86] I. C. Sluimer, A. Schilham, M. Prokop, and B. van Ginneken. Computer analysis of computed tomography scans of the lung: a survey. *IEEE Transactions on Medical Imaging*, 25(4):385–405, 2006.

[87] I. C. Sluimer, P. F. van Waes, M. A. Viergever, and B. van Ginneken. Computer-aided diagnosis in high resolution CT of the lungs. *Medical Physics*, 30(12):3081–3090, 2003.

[88] K. Soejima, K. Yamaguchi, E. Kohda, K. Takeshita, Y. Ito, H. Mastubara, T. Oguma, T. Inoue, Y. Okubo, K. Amakawa, H. Tateno, and T. Shiomi. Longitudinal follow-up study of smoking-induced lung density changes by high-resolution computed tomography. *American Journal of Respiratory and Critical Care Medicine*, 161(4):1264–1273, 2000.

[89] L. Sørensen, P. Lo, H. Ashraf, J. Sporring, M. Nielsen, and M. de Bruijne. Learning COPD sensitive filters in pulmonary CT. In G.-Z. Yang, D. J. Hawkes, D. Rueckert, J. A. Noble, and C. J. Taylor, editors, *Medical Image Computing and Computer Assisted Intervention*, volume 5761 of *Lecture Notes in Computer Science*, pages 699–706. Springer, 2009.

[90] L. Sørensen, M. Loog, P. Lo, H. Ashraf, A. Dirksen, R. P. W. Duin, and M. de Bruijne. Image dissimilarity-based quantification of lung disease from CT. In Tianzi Jiang, Nassir Navab, Josien P. W. Pluim, and Max A. Viergever, editors, *Medical Image Computing and Computer Assisted Intervention*, volume 6361 of *Lecture Notes in Computer Science*, pages 37–44. Springer, 2010.

[91] L. Sørensen, S. B. Shaker, and M. de Bruijne. Texture based emphysema quantification in lung CT. In M. Brown, M. de Bruijne, B. van Ginneken, A. Kiraly, J.-M. Kuhnigk, C. Lorenz, K. Mori, and J. Reinhardt, editors, *Proc. of The First International Workshop on Pulmonary Image Analysis*, pages 5–14, 2008.

[92] L Sørensen, S. B. Shaker, and M de Bruijne. Texture classification in lung CT using local binary patterns. In D. N. Metaxas, L. Axel, G. Fichtinger, and G. Székely, editors, *Medical Image Computing and Computer Assisted Intervention*, volume 5241 of *Lecture Notes in Computer Science*, pages 934–941. Springer, 2008.

[93] L. Sørensen, S. B. Shaker, and M. de Bruijne. Quantitative analysis of pulmonary emphysema using local binary patterns. *IEEE Transactions on Medical Imaging*, 29(2):559–569, 2010.

[94] T. Stavngaard, S. B. Shaker, K. S. Bach, B. C. Stoel, and A. Dirksen. Quantitative assessment of regional emphysema distribution in patients with chronic obstructive pulmonary disease (COPD). *Acta Radiologica*, 47(9):914–921, 2006.

[95] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[96] Q. Tao, S. D. Scott, N. V. Vinodchandran, T. T. Osugi, and B. Mueller. Kernels for generalized multiple-instance learning. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 30(12):2084–2098, 2008.

[97] B. M. ter Haar Romeny. *Gaussian Scale-Space Theory*, chapter Applications of scale-space theory, pages 3–19. Dordrecht: Kluwer Academic Publishers, 1997.

[98] M. W. Trosset, C. E. Priebe, Y. Park, and M. I. Miller. Semisupervised learning from dissimilarity data. *Computational Statistics and Data Analysis*, 52(10):4643–4657, 2008.

[99] M. Tuceryan and A. K. Jain. *The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, chapter Texture Analysis, pages 207–248. World Scientific Publishing, 1998.

[100] Y. Uchiyama, S. Katsuragawa, H. Abe, J. Shiraishi, F. Li, Q. Li, C.-T. Zhang, K. Suzuki, and K. Doi. Quantitative computerized analysis of diffuse lung disease in high-resolution computed tomography. *Medical Physics*, 30(9):2440–2454, 2003.

[101] D. Unay, A. Ekin, M. Cetin, R. Jasinschi, and A. Ercil. Robustness of local binary patterns in brain MR image analysis. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2098–2101, 2007.

[102] R. Uppaluri, E. A. Hoffman, M. Sonka, P. G. Hartley, G. W. Hunninghake, and G. McLennan. Computer recognition of regional lung disease patterns. *American Journal of Respiratory and Critical Care Medicine*, 160(2):648–654, 1999.

[103] R. Uppaluri, T. Mitsa, M. Sonka, E. A. Hoffman, and G. McLennan. Quantification of pulmonary emphysema from lung computed tomography images. *American Journal of Respiratory and Critical Care Medicine*, 156(1):248–254, 1997.

[104] B. van Ginneken, L. Hogeweg, and M. Prokop. Computer-aided diagnosis in chest radiography: beyond nodules. *European Journal of Radiology*, 72(2):226–230, 2009.

[105] B. van Ginneken, S. Katsuragawa, B. M. ter Haar Romeny, K. Doi, and M. A. Viergever. Automatic detection of abnormalities in chest radiographs using local texture analysis. *IEEE Transactions on Medical Imaging*, 21(2):139–149, 2002.

[106] J. van Sickle. Analyzing correlations between stream and watershed attributes. *Journal of the American Water Resources Association*, 39(3):717–726, 2003. Errata: 41(3) 741–741.

[107] J. Wang and J.-D. Zucker. Solving the multiple-instance problem: A lazy learning approach. In P. Langley, editor, *International Conference on Machine Learning*, pages 1119–1126. Morgan Kaufmann, 2000.

[108] W. R. Webb, N. L. Müller, and D. P. Naidich. *High-Resolution CT of the Lung*. Lippincott Williams & Wilkins, third edition edition, 2001.

[109] X. Wei. Gray level run length matrix toolbox v1.0. Software, Beijing Aeronautical Technology Research Center, 2007.

[110] D. Wu, J. Bi, and K. Boyer. A min-max framework of cascaded classifier with multiple instance learning for computer aided diagnosis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1359–1366. IEEE Computer Society Press, 2009.

[111] Y. Xu, M. Sonka, G. McLennan, J. Guo, and E. A. Hoffman. MDCT-based 3-D texture classification of emphysema and early smoking related lung pathologies. *IEEE Transactions on Medical Imaging*, 25(4):464–475, 2006.

[112] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

[113] Z.-H. Zhou. Multi-instance learning: a survey. Technical report, AI Lab, Department of Computer Science & Technology, Nanjing University, Nanjing, China, 2004.

[114] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li. Multi-instance learning by treating instances as non-i.i.d. samples. In A. P. Danyluk, L. Bottou, and M. L. Littman, editors, *International Conference on Machine Learning*, volume 382 of *ACM International Conference Proceeding Series*, pages 1249–1256. ACM, 2009.

List of Publications

Papers in international journals

- L. Sørensen, M. Nielsen, P. Lo, H. Ashraf, J. J. H. Pedersen, and M. de Bruijne, "Texture-Based Analysis of COPD: a Data-Driven Approach," submitted, 2010.

- L. Sørensen, S. B. Shaker, and M. de Bruijne, "Quantitative Analysis of Pulmonary Emphysema Using Local Binary Patterns," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 559–569, 2010.

Papers in conference proceedings

- L. Sørensen, M. Loog, P. Lo, H. Ashraf, A. Dirksen, R. P. W. Duin, and M. de Bruijne, "Image Dissimilarity-Based Quantification of Lung Disease from CT," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 6361, pp. 37–44, 2010.

- M. J. Gangeh, L. Sørensen, S. B. Shaker, M. S. Kamel, M. de Bruijne, and M. Loog, "A Texton-Based Approach for the Classification of Lung Parenchyma in CT Images," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 6363, pp. 595–602, 2010.

- M. J. Gangeh, L. Sørensen, S. B. Shaker, M. S. Kamel, and M. de Bruijne, "Multiple Classifier Systems in Texton-Based Approach for the Classification of CT Images of Lung," in *Medical Computer Vision 2010: Recognition Techniques and Applications in Medical Imaging*, ser. Lecture Notes in Computer Science, pp. 31–41, 2010.

- L. Sørensen, M. Loog, D. M. J. Tax, W.-J. Lee, M. de Bruijne, and R. P. W. Duin, "Dissimilarity-Based Multiple Instance Learning," in *Structural, Syntactic, and Statistical Pattern Recognition*, ser. Lecture Notes in Computer Science, vol. 6218, pp. 129–138, 2010.

- L. Sørensen, P. Lo, H. Ashraf, J. Sporring, M. Nielsen, and M. de Bruijne, "Learning COPD Sensitive Filters in Pulmonary CT," in *Medical Image Com-*

puting and Computer-Assisted Intervention, ser. Lecture Notes in Computer Science, vol. 5762, pp. 699–706, 2009.

- L. Sørensen and M. de Bruijne, "Dissimilarity Representations in Lung Parenchyma Classification," in *SPIE Medical Imaging*, vol. 7260, pp. 72602Z1-72602Z12, 2009.
- L. Sørensen, S. B. Shaker, and M. de Bruijne, "Texture Based Emphysema Quantification in Lung CT," in *The First International Workshop on Pulmonary Image Analysis*, pp. 5–14, 2008.
- L. Sørensen, S. B. Shaker, and M. de Bruijne, "Texture Classification in Lung CT Using Local Binary Patterns," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 5241, pp. 934–941, 2008.
- L. Sørensen, J. Østergaard, N. Jørgensen, P. Johansen, and M. de Bruijne, "Multi-Object Tracking of Human Spermatozoa," in *SPIE Medical Imaging*, vol. 6914, pp. 69142C1–69142C12, 2008.

Published abstracts

- L. Sørensen and M. de Bruijne, "Pattern Recognition Based Emphysema Quantification," in *The 16th Danish Conference on Pattern Recognition and Image Analysis*, DIKU Technical Report no. 08-10, pp. 4–5, 2008.

Patent applications

- L. Sørensen, M. Nielsen, and M. de Bruijne, "Classification of Medical Diagnostic Images", US patent application, filed 11 May 2010.

Acknowledgements

First and foremost, I would like to thank my supervisor Marleen de Bruijne for excellent supervision and support during the project, and from whom I have learned a lot! I would also like to thank my co-supervisor Mads Nielsen for occasional "pep-talks" and for sharing his view on the larger context in which the project fits and my co-supervisor Jon Sporring for always approaching and explaining the problems encountered in the project from a mathematical angle, providing me with a different view.

I also owe thanks to all the people in The Image Group at the Department of Computer Science, both current and former members, who all contributed to an inspiring and great working environment. Special thanks go to the two other Ph.D. students in the "lung gang", Pechin Lo and Vladlena Gorbunova, with whom I have shared many great experiences, and to my old supervisor Peter Johansen who initially drew my attention to this Ph.D. position and encouraged me to apply. Further, I would like to thank all the people from Gentofte University Hospital and AstraZeneca R&D Lund who participated in the project "Computer-Aided Assessment of COPD from CT Images", which this Ph.D. project was a part of, for an excellent cross-disciplinary collaboration.

I would also like to thank all the members and guests of the Pattern Recognition Laboratory at Delft University of Technology in The Netherlands where I stayed for five inspiring months. Especially thanks to Robert Duin, Marco Loog, and David Tax for their "open door" policies which I used heavily during my stay.

Finally, thanks to my parents and my two sisters for always supporting me, and to my beloved girlfriend Camilla – you always show genuine interest, and support me, in my work, and you often ask the toughest questions ;-).