Automatic detection of tuberculosis in chest radiographs

Laurens Edo Hogeweg

This book was typeset by LaTeX 2$_\varepsilon$.

Cover design by Laurens Hogeweg and Sjoerd Kerkstra.

Financial support for publication of this thesis was kindly provided by the department of radiology of the Radboud University Nijmegen Medical Centre (Nijmegen, the Netherlands) and Delft Imaging Systems (Veenendaal, the Netherlands).

Printed by Ipskamp Drukkers (Nijmegen, the Netherlands).
ISBN: 978-94-6191-921-2

Automatic detection of tuberculosis in chest radiographs

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S. C. J. J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op woensdag 27 november 2013
om 16.30 uur precies

door

Laurens Edo Hogeweg

geboren op 17 augustus 1982
te Hoorn

Promotor:	Prof. dr. B. van Ginneken
Co-promotor:	Dr. ir. Clara I. Sánchez
Manuscriptcommissie:	Prof. dr. ir. Nico Verdonschot Prof. dr. ir. Bart ter Haar Romeny Dr. Martin Boeree

The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Nijmegen Medical Centre (Nijmegen, the Netherlands) and the Image Sciences Institute, University Medical Center Utrecht (Utrecht, the Netherlands).

This work was funded in parts by SenterNovem project Computer Aided Diagnosis for Tuberculosis (CAD4TB) of the Dutch Ministry of Economic Affairs, Agriculture and Innovation; the Tuberculosis Diagnostic Platform (TBDX) project of the Dutch Ministry of Economic Affairs, Agriculture and Innovation; and the Evaluation of multiple novel and emerging technologies for TB diagnosis, in smear-negative and HIV-infected persons, in high burden countries (TB-NEAT) project of the European and Developing Countries Clinical Trials Partnership (EDCTP).

TABLE OF CONTENTS

1 Introduction **1**
1.1 Tuberculosis . . . 2
1.2 Chest radiography . . . 7
1.3 Computer aided detection (CAD) . . . 12
1.4 Automatic analysis of chest radiographs . . . 20
1.5 Thesis outline . . . 23

2 Foreign object detection and removal **25**
2.1 Introduction . . . 27
2.2 Data . . . 29
2.3 Methods . . . 31
2.4 Experiments . . . 38
2.5 Results . . . 41
2.6 Discussion . . . 49
2.7 Conclusion . . . 54

3 Clavicle segmentation **55**
3.1 Introduction . . . 57
3.2 Data . . . 59
3.3 Methods . . . 61
3.4 Experimental . . . 68
3.5 Results . . . 69
3.6 Discussion . . . 75
3.7 Conclusion . . . 81

4 Suppression of translucent elongated structures **83**
4.1 Introduction . . . 85
4.2 Methods . . . 87
4.3 Experiments & Results . . . 93
4.4 Discussion . . . 107
4.5 Conclusion . . . 113

5 Quantification of symmetry **115**
5.1 Introduction . . . 117

5.2 Methods . . . 119
5.3 Experiments & Results . . . 125
5.4 Discussion . . . 135
5.5 Conclusion . . . 139

6 Automatic detection of tuberculosis 141
6.1 Introduction . . . 143
6.2 Methods . . . 145
6.3 Materials . . . 152
6.4 Experiments & results . . . 153
6.5 Discussion . . . 162
6.6 Conclusion . . . 166
6.A Training and testing of CAD components . . . 167
6.B Results for all combination methods . . . 168

7 Diagnostic accuracy of the automated system 171
7.1 Introduction . . . 173
7.2 Methods . . . 174
7.3 Results . . . 178
7.4 Discussion . . . 182
7.5 Conclusion . . . 184
7.A Methods supplement . . . 185

Summary and discussion 191

Samenvatting 207

Publications 213

Bibliography 217

Acknowledgements 239

Curriculum Vitae 243

Introduction

1

Tuberculosis is a common disease with high morbidity and mortality rates worldwide. Chest radiography plays an important role in screening algorithms. The introduction of digital radiography has made it easier to develop automated systems that detect abnormalities related to tuberculosis in chest radiographs. This thesis describes the development of such an automated system and its evaluation. This thesis is part of the larger Computer Aided Detection for Tuberculosis (CAD4TB) project*.

1.1 Tuberculosis

Despite the existence of an effective and affordable cure, tuberculosis (TB†) remains one of the world's major health care challenges. Mortality and morbidity rates are only slightly lower than those of the well known HIV/AIDS epidemic[1], but TB has received less attention of the media and public. One of the reasons for this has been the decline of TB in high-income countries[2]. In recent years attention for TB has increased again; novel diagnostics and drugs are being developed, but there are also pressing challenges on the horizon with the emergence of multi-drug resistant (MDR) and extensively drug resistant (XDR) TB[3].

1.1.1 Epidemiology

In 2011 an estimated 8.7 million new cases and 1.4 million deaths were reported[1]. The majority of the TB burden is located in the low and middle income countries[1], although an increase of cases has been reported in selected populations in (several of) the more affluent parts of the world[4,5]. In terms of incidence, the number of new cases per population unit, Sub-Saharan Africa is most affected with rates going as high as 800 cases per 100,000 in some countries[1]. The determining reason for the high rates in this area is co-infection with HIV. In absolute numbers most cases are found in Asia, especially India and China. There is a trend of decreasing global incidence since 2004, although the total number of cases has increased as a result of the growth of the world population[2]. About two billion people (roughly one third of the world population) have latent TB infection. They do not have the active form of disease, but carry the bacteria in a dormant state. While most of these people never progress to active disease, the progression can happen whenever their immune system is compromised.

*`http://www.diagnijmegen.nl/index.php/CAD4TB`

†TB is short for *tubercle bacillus*, the common abbreviation for tuberculosis

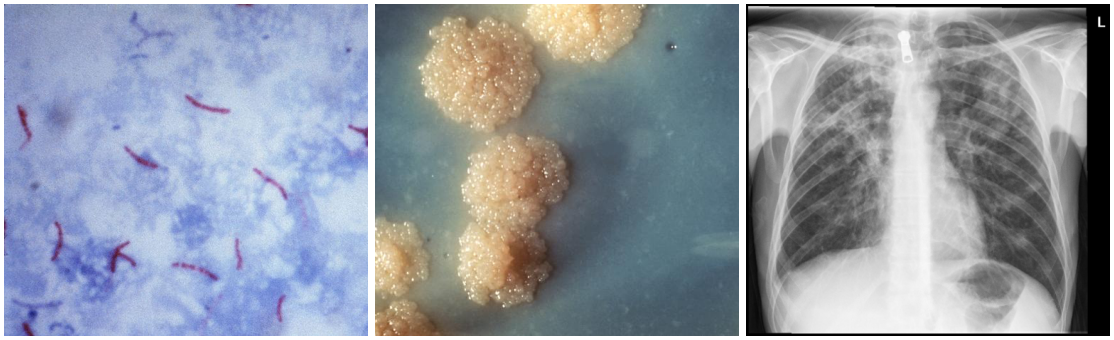
1.1.2 Pathogenesis

Although TB can affect almost any part of the body, its main site of infection is the lungs because of the bacillus' preference for high oxygen environments. The pathogenesis of TB is complex, and some of its features are not fully understood yet. Bacilli enter the lungs through the airways and end up in the alveoli where they invoke the innate immune response. Macrophages ingest the bacillus and try to destroy it. If the host is unable to do that, for example when it has not encountered TB before, the bacillus will replicate, destroy the macrophage and release its many copies. This process will continue until the acquired immune system eventually contains the infection. In the majority of cases this process is sufficient to heal the patient, who, being asymptomatic, remains unaware of the infection. Only approximately 5% of infections lead to clinically active disease[6].

Symptoms of the disease are caused by the immune response and the destruction of lung tissue in the process[7]. The immune response causes inflammation, leading to a local increase of density in the lung tissue. If the disease remains at the site of infection, monocytes will attempt to wall off the hazard leading to formation of a granuloma, called *tubercle* in TB. In some cases the granuloma progresses and the monocytes release a number of chemicals that destroy bacilli and lung tissue alike, leading to so-called caseous necrosis. The hallmark sign of TB, the cavity, is a result of this process, although it does not occur in all cases[8]. When the bacilli spread through the lymphatic system they can infect the lymph nodes in the hilar and mediastinal area, leading to lymphadenopathy. This type of lesion is more common in childhood TB, where it is also often the only sign. Bacilli that enter the pleural space can cause an increase of pleural fluid, leading to effusions. When TB is not contained to the lung tissue and enters the bloodstream it can infect large parts of the lung hematogenously and lead to miliary TB. This manifestation is uncommon in subjects with a normal functioning immune system.

1.1.3 Diagnosis

In clinical practice, TB is diagnosed using a combination of clinical symptoms, chest radiography, and sputum examination (Fig. 1.1). The typical symptoms associated with TB are fever, weight loss, night sweats, and coughing. However, it is also possible for people with active disease, in which the bacilli are present in the sputum and thus are able to infect others, to present without any of these symptoms. These asymptomatic patients are diagnosed with chest radiography



(a) Sputum smear microscopy (b) Sputum culture (c) Chest radiography

Figure 1.1: Traditional diagnostic techniques for TB

and tests examining the sputum. The most commonly used test in high burden areas is sputum smear microscopy. This test involves coughing up sputum by the patient, which is then fixated and stained on a slide. The presence of the characteristic TB bacillus indicates a positive test (Fig. 1.1(a)). Smear microscopy has a high specificity but a relatively low sensitivity[9]. The introduction of LED fluorescence microscopy has improved sensitivity compared to conventional microscopy and is recommended by the World Health Organization (WHO)[9]. In populations with high HIV incidence, the already low sensitivity of smear microscopy is reduced even further[10]. Therefore, the WHO recommends the use of chest radiography in high HIV prevalence populations[11]. The gold standard for diagnosis of TB is the sputum culture test (Fig. 1.1(b)), which has both a high sensitivity and specificity. Sputum culture has as disadvantages that it can take up to six weeks before the result is known, it is relatively expensive, and the method requires good laboratory facilities. The unavailability of a good point-of-care* test for TB has led to the development of numerous novel diagnostics[12]. Nucleic acids amplification tests are the most promising, and one of them, GeneXpert, has good sensitivity and specificity and has shown first promising results[13]. This test can also determine resistance of the bacillus to rifampicin, one of the most important TB drugs. The main drawback for the test is its currently high cost of 10$ per test compared to 4$ and 2$ for sputum smear and chest radiography, respectively. Sputum culture costs about 20$ per test. Note that these numbers are estimates as reported costs in literature vary substantially. Also, cost-effectiveness analysis concerns more than the cost per test. Other factors, such as the test's perfor-

*The definition of point-of-care is contentious, but, at the least, it implies the ability to make a diagnosis at the point where patient consultation and presentation occurs, and the ability to translate the result into same day treatment, if appropriate.

	Setting	Reference standard	Sensitivity	Specificity	Type of readers	Analog (A)/ Digital (D)
van Cleef et al.[15]	Chest clinic	Bacteriological	91	67	Radiologists	A
den Boon et al.[16]	Prevalence survey	Bacteriological	95	67	Pulmonologist	A
Lewis et al.[17]	Miners	Bacteriological	26	99	Physicians	-
Dawson et al.[18]	HIV infected	Bacteriological	68	53	Radiographers	A
van 't Hoog et al.[19]	Prevalence survey	Bacteriological	94	73	Clinical officers	A
Story et al.[20]	High-risk group screening	Bacteriological and clinical	82	99	Radiographers	D
van 't Hoog et al.[21]	Prevalence survey	Bacteriological	82	76	Experts	A

Table 1.1: Selected literature concerning CXR performance for TB detection.

mance, required infrastructure, prevalence and personnel costs, must be taken into account to obtain a good estimate of the cost per detected TB case[14].

The role of chest radiography in the diagnostic process has been marginalized for some time by the WHO. The reason for this was a number of studies in which a large interobserver disagreement was found[22]. One of the reasons for this disagreement was the poor quality of analog chest radiographs (CXRs). A problem which was addressed by the introduction of digital radiography (Fig. 1.1(c) and Section 1.2). Like other diagnostic tests, chest radiography has a number of limitations. Most studies report high sensitivities, but low specificities[21]. This is a consequence of abnormalities on the CXR not being specific to TB, as they can occur in other diseases as well. As can be seen in Table 1.1 varying pairs of sensitivity and specificity are reported. The different performances can be a result of the specific population and setting in which the study was performed, but are also due to the subjective nature of CXR reading. Each reader may have a different threshold for calling abnormalities sufficiently suspicious to consider the CXR abnormal. This thesis aims at providing a objective repeatable TB suspiciousness score for CXRs based on automatic analysis of the radiograph by a computer.

1.1.4 Treatment

TB is a curable disease and diagnosing people who need treatment is essential to reduce the burden of the disease. Treatment for TB consists of a regimen of a daily cocktail of four drugs during six months. Cure rates of 90% are observed in TB control programs[1,3]. It is critical that people complete their treatment fully for a successful outcome and to prevent development of drug resistance. Therefore the directly observed treatment, short course (DOTS) strategy was initiated by the WHO in 1993, leading to treatment of 43 million people in the period from

1995 till 2008 and a reduction of fatal cases[2]. Both HIV infection and multidrug resistant TB require other more expensive treatment strategies[2].

1.1.5 Types of TB detection programmes

The automated system for TB detection in CXR presented in this thesis can be integrated into programmes that are aimed at TB detection. Four types of such programmes are discussed here: high-risk group screening, mass screening, prevalence surveys, and use in a clinical setting.

High-risk group screening is related to the concept of active case finding, defined as "... looking systematically for cases of active TB and latent infection in groups known, or thought to be, at higher risk of TB, rather than waiting for people to develop symptoms/signs of active disease and present themselves for medical attention (passive case finding)."[23]. Target locations for high-risk screening for TB include prisons[24], ports of entry[25], and risk-groups in large cities[5]. In this thesis the automated system has been evaluated in high-risk groups in London[5,20]. Mass screening has the same aim as active case finding, with the difference that is not specifically directed at high-risk groups but at the whole population*. Mass screening for TB using chest radiography was common until the 1960s[26] and still is in a number of predominantly Asian countries[27]. Prevalence surveys are used to measure the burden of TB and the effect of control programs. They have been recommended to be used in 21 countries with a high TB incidence by the WHO[28]. In a clinical setting people who present with clinical symptoms suspicious of TB infection are tested. A database of TB suspects was used for evaluation in Chapter 5 and 6.

For the purpose of this thesis, the main difference between the types of programmes is the total number of evaluated cases and the ratio between normal and abnormal cases. In mass screening and prevalence surveys the number of abnormal cases is low and the total numbers are very high. In high-risk group screening the prevalence is also low, but typically higher than in mass screening. The total number of cases is lower than in mass screening or prevalence surveys. In a clinical setting and in endemic countries the prevalence is usually much higher than in screening programs and the total number of cases lower. The large number of screened people in prevalence surveys and screening programs prohibit the use of confirmatory diagnostic tests on everybody. Therefore a strategy consisting of a cascade of tests is often used, with the aim of selecting

*Some authors include mass screening among active case finding strategies[26].

cases that require further testing. The optimal strategy depends on the diagnostic performance of the individual tests in the specific population that is screened. A number of such strategies were evaluated by van 't Hoog et al.[21]. They concluded that a combination of symptom and CXR screening gives the best results. We propose to include the automated system described in this thesis as one of these tests. The varying requirements of TB detection programmes require that the automated system can be adapted to them. A continuous TB likelihood score and a system that can learn the differences between populations is presented in Chapter 6. An important factor in such studies is the cost per detected case of each of the screening strategies; a cost-reduction or improvement in diagnostic performance may change the optimal strategy[29]. The use of digital radiography and reduction of cases requiring human evaluation (described in Chapter 7) may contribute to the goal of cost reduction.

1.2 Chest radiography

1.2.1 History

The ability to look inside parts of the living human body that were previously inaccessible started with the discovery of X-rays by Wilhelm Röntgen in 1895[30]. Almost immediately after the discovery the first radiographs of hands were taken. The chest radiograph quickly followed and has been the work horse of radiology since, in terms of number of exams acquired per year[31]. Its visualization of the critical organs in the thorax have provided radiologists with a wealth of diagnostic information[32,33].

Contrast in the CXR is achieved through different densities of anatomical structures and thus varying X-ray absorption rates. Different amounts of X-ray radiation hit the imaging plane (historically film, now often a digital detector) and lead to intensity differences in the image. The default convention for display on screen is that white indicates high density and black low density. Unlike computed tomography the values in the image are not absolute density measures. The standard view is when the patient faces the imaging plane, and X-rays enter on the posterior side of the patient and exit on the anterior side, giving a posterior-anterior (PA) CXR. Note that the left side of the patient is displayed on the right of the image. Anterior-posterior (AP) and lateral CXRs are acquired less regularly. In this thesis all CXRs are PA acquisitions. The spatial resolution depends on the quality of the equipment, in this thesis CXRs with resolutions

Figure 1.2: One of the first digital units that is currently operational for the CAD4TB project was installed in Kanyama Clinic, in Lusaka, Zambia. In this clinic more than 10,000 TB suspects are seen every year.

varying from 0.15 mm to 0.25 mm were used.

Digital chest radiography has led to improved image quality, obviated the use of films, chemicals, and water, and making the image directly available[34]. Films are expensive, and therefore a digital radiography unit is much cheaper than an analog unit, especially if the unit acquires a large number of images per day*. Digital chest radiograpgy also has enabled teleradiology, where images can be diagnosed remotely when local expertise is not available[35]. Another great benefit of digital radiography is that it accelerated the development of computer algorithms that analyze the image automatically, as it greatly simplifies the collection of large image databases. Digital chest radiography requires, just like conventional radiography, an electrical power source for the generation of the x-ray beam, but also for the detector plate and the computer running a digital archiving system. Robust solutions have been developed that allow the use of digital x-ray cameras in mobile vans or standard size sea containers (Fig. 1.2). These rely on batteries, generators, or solar power. Although wired Internet is not widely available in most African countries, the wireless network is sufficiently fast to send (suitably compressed[36]) digital images from anywhere, including the most rural areas, to a central point for storage or reading. In this manner a large part of the CAD4TB database that was used in this thesis was collected, with images being automatically sent from sites in Zambia and South Africa.

*See `http://www.checktb.com` for an indication of the costs of analog and digital radiography.

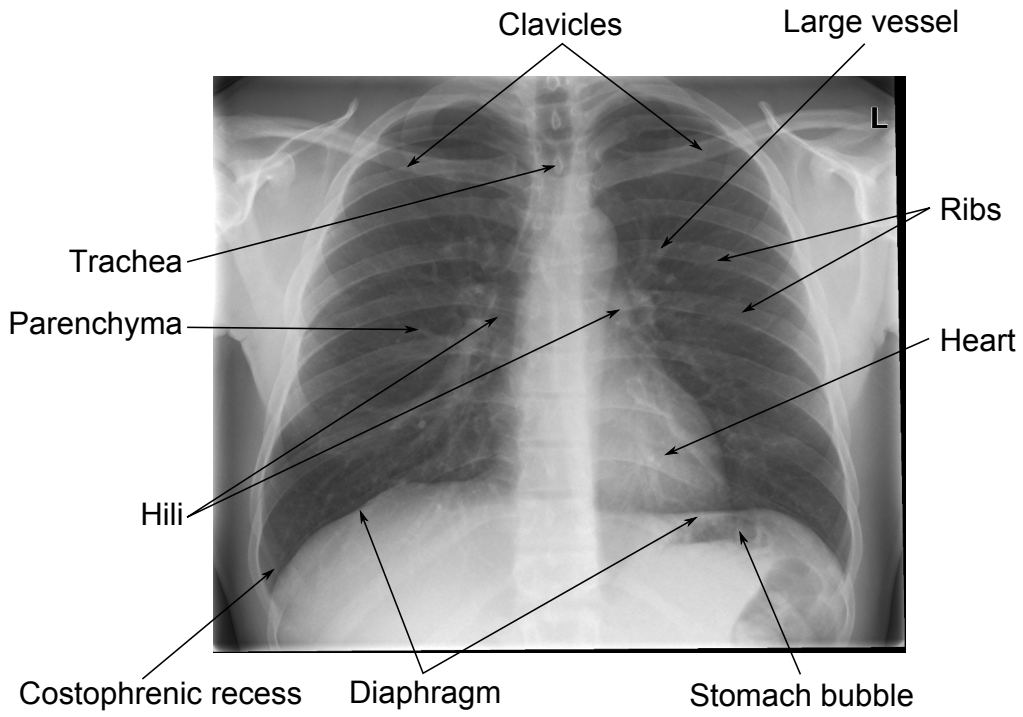
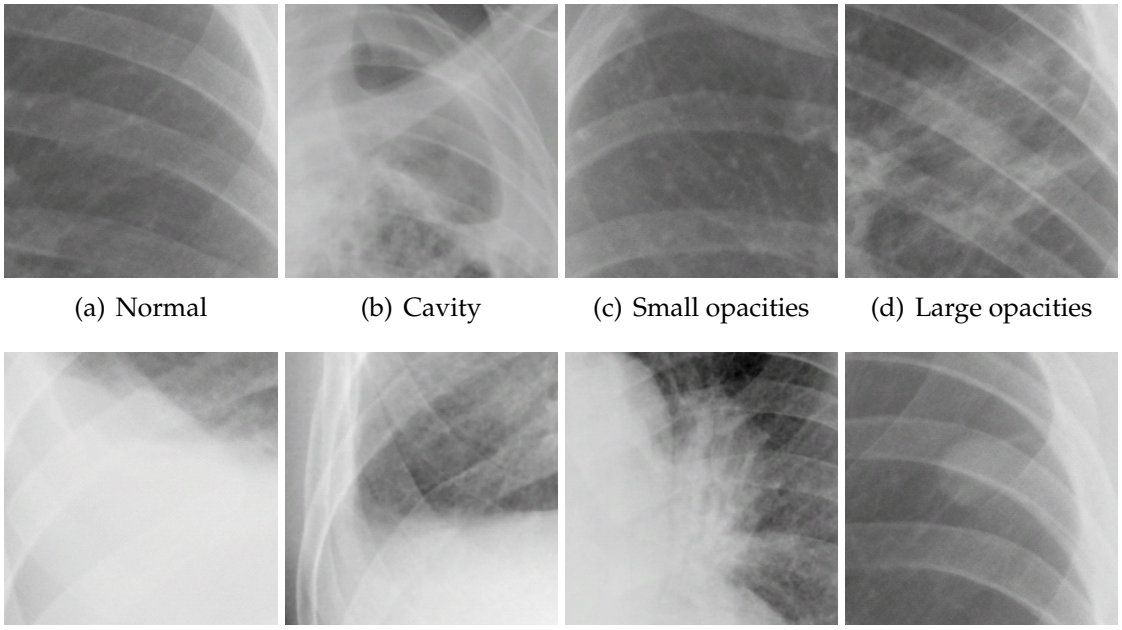


Figure 1.3: Digital radiograph with anatomical structures indicated.

1.2.2 Lung anatomy

On the CXR the lungs are demarcated by the rib cage on the lateral and superior sides, the mediastinum and heart on the medial side, and the diaphragm on the inferior side (Fig. 1.3). Because the CXR is a projection image some structures, such as the heart and diaphragm obscure part of the lungs. The low density (dark in standard display) parts, where only the ribcage overlaps, are referred to as the unobscured lung fields. Analysis in this thesis was focused on the unobscured lung fields, although pathology might be present in the obscured part[37]. The structures that can typically be easily identified in healthy unobscured lung fields are the clavicles, ribs and large vessels. Major airways are sometimes visible but in healthy patients typically only to the level where the trachea divides in the two major bronchi. The hili, where lung vessels and airways enter the lung, are visible in the medial and central part of the lung fields. The functional lung tissue, or parenchyma, is only visible as a slight, non-uniform, increase of density. Individual functional structures, such as the alveoli, can not be distinguished. The pleura, providing a smooth interface between lungs and rib cage, and the pleural space between them, are not visible on the normal radiograph.



(a) Normal (b) Cavity (c) Small opacities (d) Large opacities

(e) Consolidation (f) Blunt CPR (g) Lymphadenopathy (h) Focal lesion

Figure 1.4: Examples of different types of abnormalities and normal appearance of the lung field. CPR = costophrenic recess. See text for descriptions.

1.2.3 Tuberculosis manifestations on the chest radiograph

Manifestations of TB are not limited to a single sign on the CXR (Fig. 1.4). There are many factors that can lead to different pathological patterns, including age[38,39], ethnicity[40], immune status[41], and stage of the disease[8]. None of the patterns are fully specific for TB, as they can also be encountered in other diseases, but a combination of certain patterns can discriminate between diseases.

The most specific sign is the cavity, which appears as a low-density spherical space surrounded by a wall of high density infected tissue. Inflammation and subsequent destruction of the lung tissue lead to a variety of patterns. Multiple small foci surrounded by still healthy lung tissue lead to a pattern of dots, called small opacities in the chest radiograph recording system (CRRS)[42]. The CRRS was used throughout this thesis for categorization of abnormalities, more information is given in Sect. 1.2.4. When inflammation increases and destroys more tissue, a nonuniform diffuse pattern of increased density appears, called large opacities. Increase of fluid, either blood or pleural fluid, in spaces surrounding the lung will lead to a uniform density increase, called consolidation*. An in-

*The term consolidation is sometimes used for any density increase; in this thesis it is only used for a uniform increase.

crease of pleural fluid can also manifest itself by displacement of the lung boundary, either visible as a blunting of the costophrenic recess or the appearance of an extra line near the rib cage. Infection of the lymph nodes (lymphadenopathy) leads to several signs, such as a widening of the mediastinum, shift of the mediastinum, or increased size of the hilar structures. Isolated focal lesions also occur in TB, most often in the early stages of the disease.

1.2.4 Interpretation

Before the invention of computed tomography in 1967 and magnetic resonance imaging in 1973, the CXR was the only way to image lungs *in vivo* completely. A large amount of information can be extracted from the CXR, not only about the lungs, but also the cardiovascular status and the bones (e.g. osteoporosis). Nowadays a radiologist in doubt about signs on the CXR will order another more detailed exam, usually a computed tomography (CT) scan. This has led to the interpretation of CXRs for diffuse infiltrates, a particularly difficult type of abnormality, being described as a dying art[43]. Nevertheless, CXR reading is still a basic part of the radiologist training.

Methods that teach how to read and interpret the CXR typically propose a form of structured reading, in which all the aspects of the CXR are systematically evaluated. Experienced chest radiologists typically integrate this systematic reading into a more holistic approach in which many abnormalities are quickly detected in a form of pattern recognition, after which a scrutiny of all the suspect areas follows[44]. The novice CXR reader encounters a number of pitfalls: superimposition of structures can be mistaken for abnormalities, the appearance can differ substantially also in the normal CXR, and subtle abnormalities can easily hide in the complex pattern of other structures. In tasks that require an objectification of findings, such as quantification of one's certainty about the finding being an expression of disease, it is known that considerable disagreement between readers can arise. This has led to the development of scoring systems, which require the reader to assign numerical values to the presence of carefully defined categories of findings. A well-known scoring system for CXRs is the international labor office (ILO) classification, which was developed to objectively judge the presence of pneumoconiosis (a disease affecting amongst others coal miners)[45*]. The ILO classification system was modified specifically for TB research in South

*There is also a political aspect to the ILO scoring system: in some countries, financial compensation for disease, as a consequence of having worked in the mining industry, depends on the outcome of the classification.

Africa. This system, the chest radiograph recording system (CRRS), has seen several evolutions[18,42]. One of the motivations behind the system is that judgment of the CXR should be based on distinguishable visual patterns, without directly giving an interpretation in terms of pathology. This makes the system suitable for readers who are not medically trained. The visual categories of the CRRS were used in this thesis. An overview of other scoring systems for CXR is given in Pinto et al.[46]. While scoring systems can reduce disagreement between readers, the assignment of numerical values is still subjective. One of the longer term aims of the CAD4TB project is to replace manual scoring by repeatable automatic computerized scoring.

1.3 Computer aided detection (CAD)

The automatic analysis of medical images is a broad field where the disciplines of medicine and computer science meet. Even though strictly taken computer aided detection (CAD) refers to software aiding the radiologist in the detection of normal or abnormal structures[47], we prefer the broader definition given in Giger et al.[48] as: "the use of computer algorithms to aid the image interpretation process". CAD as a field started in the 1960s, with a method to analyze pulmonary lesions in CXRs[49]. Since then, fueled by increasing computing power and increasing availability of digital image databases, the number of publications has quickly increased[50].

1.3.1 Design of CAD systems

Pioneering CAD researchers were very ambitious in their aim to replace humans by automatic reading. In that, they underestimated the complexity and number of required steps involved in analyzing an image for abnormalities. The development of a CAD system is therefore typically broken into a number of discrete steps or components which can be addressed individually. These steps are often based on how radiologists work, or at least how they are thought to do it, because it is often very difficult for experts to translate their knowledge into discrete procedures that computers can follow. From a more mathematical point of view a CAD system can be summarized as a reduction of information, by discarding what is irrelevant to diagnosis. In this thesis, the CAD system often produces only one number, though it started with millions of numbers (pixel values) per image.

We distinguish five basic phases in the operation of a CAD system: preprocessing, segmentation, feature calculation, classification, and combination. The last phase, combination, is not typically mentioned explicitly in the design of CAD systems. A special emphasis is attributed to this phase because it is an important concept in this thesis and we believe it is one of the ways forward in CAD research. In concrete instances of CAD systems these components are often arranged in a complex network, instead of in a simple serial execution of the phases.

Preprocessing

The goal of preprocessing is twofold. First, it serves the purpose of reducing differences between images, that are not the result of differences between patients, but of differences in acquisition technique. One can think of different brands of imaging devices, different settings used during acquisition, and operator dependent patient positioning. For radiologists these differences sometimes pose a problem, but they have in general a large ability to ignore variations that are irrelevant to the task at hand. On the contrary, computers do not have this ability, and have to be made explicitly aware of these differences. General techniques for removing differences include histogram equalization, scaling, and frequency spectrum normalization. The second reason for preprocessing images is to enhance or decrease the visibility of certain structures in the image. These structures can be very basic, such as edges, or more geared towards the application, such as an enhancement of spherical objects to detect nodules in lung images. In this thesis preprocessing was used to remove ribs, clavicles, catheters, and other foreign objects from the image.

Segmentation

Segmentation involves the division of the image into several areas. The areas can be based on their distinct visual properties, but in medical imaging the goal of segmentation is usually to outline specific anatomical structures. Knowing where these structures are provides the other computations with a coordinate framework, which is not based on the abstract distances in the image grid, but the content of the image. Two types of segmentations can be discerned; the binary segmentation in which a pixel in the image is assigned to a single structure at the time, or a probabilistic segmentation in which a likelihood of belonging to different structures is assigned. This distinction is of importance because, as all algorithms in image analysis, segmentation is prone to make errors, and there is

valuable information contained in the likelihood values. This information was used to improve clavicle segmentation in Chapter 3.

Features

Many CAD algorithms described in literature, including most of the methods described in this thesis, follow a supervised pattern recognition approach[51]. The basis of this approach is the representation of the objects of interest (pixels, regions, images) as a vector of numerical characteristics describing their properties. One characteristic is also called a *feature*, and the vector then the features of the object. The vectors of all the objects span a multidimensional space, the *feature space*, in which classification is performed. Any characteristic which contains information that is useful to perform the task can be used as a feature. Most features try to provide a concise description of the data, fitting in the "CAD as reduction" paradigm. An example is the description of an image region by the average pixel value.

There are two basic approaches to define image features. The first tries to translate a specific visual property, for example indicated by a radiologist as important, into an algorithm which can be used in an image consisting of pixels. An illustration of this process is the translation of the visual pattern that describes the invasion of breast tumor masses in the surrounding tissue in a spiculation feature[52]. The other approach uses a set of more general features to describe a pixel or region. These methods are sometimes inspired on the workings of the human visual system in the brain. Texture features are an example of such features, describing the local structure and appearance of the image in terms of the distribution of values[53]. This type of features plays an important role in the analysis of CXRs for TB. These features have not been designed to describe specific image properties, e.g. related to a disease, and leave the determination of their relevance to the classification step. Both approaches can be included in the design phase of a specific application.

Information is usually lost in the process of feature computation, but this is essential for most classification methods to be able to operate in a practically usable timeframe. It is also required because of the curse of dimensionality, a concept from pattern recognition in which larger amounts of training data are required in higher dimensional spaces. Therefore, much attention has been focused on methods that reduce the number of features by selecting the most interesting ones, either through supervised feature selection or unsupervised dimensionality re-

duction approaches. A well-known example of the latter is principal component analysis (PCA) and its many related techniques. PCA is used throughout this thesis for that purpose.

Classification

The basic idea of classification is to divide a set of objects into multiple classes based on a rule or procedure derived from the characteristics/features of the objects. This process is also referred to as assigning labels to objects. For example, a goal can be to assign to a set of CXRs the labels ”normal”, ”active TB”, or ”other disease”. In general the number of classes is not restricted, but often the problem is formulated as a two class classification. Many machine learning techniques have been developed specifically for two class problems[54]. There is also an intuitive procedure available to reduce any multi-class (more than two classes) problem to a set of two class problems[55]. In this thesis most of the classification problems are of the two class kind.

An import concept in classification is that of class overlap. Class overlap means that given the features it is not possible to find a rule which perfectly separates the objects into the two classes in the feature space. Because of this, objects will end up on the wrong side of the decision boundary and be assigned the wrong label. The decision boundary is a hypersurface which partitions the feature space into regions for different classes. Multiple complex hypersurfaces can exist in the same feature space. The objects on the wrong side of the decision boundary are classification errors.

A procedure that implements a specific classification algorithm is called a classifier. Many types of classifiers exist, sometimes they are based on extensive theory and complex procedures and sometimes they are intuitive to understand[51,54]. Of the parametric classifiers, which assume a Gaussian distribution of the data, linear and quadratic discriminant analysis are well-known examples. Linear discriminant analysis is known for its speed and occasionally good results and was used in Chapters 2, 3, 6, and 7 for various purposes, although often for comparison with other classifiers. The support vector machine[56] which searches for a decision boundary with the largest margin between the two classes, is a popular and high-performing classifier, although it is computationally expensive. It was used in Chapter 2 for the detection of foreign objects. k-nearest-neighbor is an intuitively simple classifier and often has good results, but can be slow in the testing phase. It was used for detecting clavicles and foreign objects in Chapters 2 and

3, respectively. Recently the family of ensemble classifiers, which intelligently combine a number of other classifiers, has become popular, due to their good performance and resilience against overfitting. Overfitting is a problem of some classifiers, in which the classification procedure works well for training data but not for test data. They have appealing names like AdaBoost[57], GentleBoost[58] or Random Forest[59]. These types of classifiers where used in Chapters 2, 6, and 7 for various tasks.

Most classifiers require a training phase in which its parameters are learned from the data. In the test phase unseen objects are then classified. It is important that the labels of the unseen objects are hidden from the training procedure as the goal is to find a procedure which works for unseen and unlabeled objects. Strictly taken classification means that objects are assigned to one label only, so-called hard classification. In practice most of the classifiers and applications provide soft classification, where a number is assigned for each of the classes, which reflects its likelihood of belonging to a label. Some classifiers, such as linear discriminant analysis or k-nearest-neighbor classification, provide estimates of true probabilities. For most applications, such as performance evaluation, it is only required that the classifier imposes an ordering on the objects, based on the resemblance to a label.

In unsupervised classification the objects are grouped into classes based on similarities between the features. No label information is used and therefore no training phase is required. Because in CAD the goal is to detect specific types of normal or abnormal structures, unsupervised classification on its own is not often used. It is useful though as a part of other components or to explore the data during the design phase of the CAD system. A well-known method is clustering, using for example the k-means algorithm, which was used in Chapter 4 to separate outlying structures from the structures of interest for the suppression of elongated structures. Feature reduction techniques are also sometimes grouped under the umbrella of unsupervised methods. They were used in Chapter 2 to discover relevant features for detecting foreign objects.

Combination

Combination of information is one of the key concepts in CAD development and machine learning in general. Implicitly, combination is present in many of the components of the CAD system. For example, a classifier combines the different features describing the object into one number. Where combination becomes ex-

(a)

TB \ CXR	Positive	Negative
Positive	TP	FN
Negative	FP	TN

(b)

TB \ CXR	Positive	Negative	Total
Positive	64	23	87
Negative	11	102	113
Total	75	125	200

Table 1.2: Confusion matrices. (a) Definition of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). (b) Confusion matrix for human CXR reading and TB diagnosis by sputum culture (example from Chapter 6).

plicit is when multiple components that could work on their own are combined to improve the performance. The basic idea of combination is that the individual combined systems produce different kinds of errors. When these errors are complementary they cancel out each other after combination[60]. Examples of this are the use of averaging, such as blurring in images, to reduce noise, and the use of voting to come to a decision. The reason that combination as concept is stressed here is that recently large improvements have been found after combination of several independently developed CAD systems[61,62]. We believe that one of the ways forward in CAD development is to focus more on combining existing techniques rather than only developing new techniques. Combination has an explicit role in the chapters on clavicle segmentation (Chapter 3) and the description of the complete CAD system (Chapter 6).

1.3.2 Evaluation

An important aspect of any procedure that classifies objects, whether it be CXRs as normal or abnormal or pixels as lung or background, is the number of errors it makes. The confusion matrix is a convenient way of summarizing the errors and forms the basis for many evaluation measures. Given a dataset of objects with true labels and assigned labels, the cells in the matrix count how many instances there are of a particular combination of true label and assigned label. For two class problems it is customary to use the general labels of being ”positive” or ”negative” for some property. For example, a patient can have active TB (he/she is ”positive”) but have a negative classification result. The cells of the confusion matrix then get a specific meaning: on the diagonal are the correctly classified *true positives* (TP) and *true negatives*, the other two cells count the *false positives* (FP) and *false negatives* (FN) (Table 1.2(a)). The true label is preferably determined by a gold standard test. For TB a gold standard is not available and the true label

is defined instead by a reference standard, which itself can contain errors. The aforementioned procedure is still valid, although caution must be taken during interpretation of experimental results.

Point performance measures

Classification performance is often reported in terms of the sensitivity = TP / (TP+FN): the percentage of positive objects being correctly labeled as positive, and the specificity = TN/(TN+FP): the percentage of negative objects correctly labeled as negative. Sensitivity and specificity are insensitive to the ratio of positive and negative objects in the evaluation dataset and thus say little about the total number of errors in a practical situation such as a screening program. It can therefore be more informative to report performance measures that depend on this ratio. Commonly used are the positive predictive value (PPV) = TP / (TP + FP): the percentage of objects labeled positive that are truly positive, and the negative predicate value (NPV) = TN / (TN + FN): the percentage of negative labeled objects that are truly negative. The NPV is used in Chapter 7 to report on the suitability of CAD as a screening test.

Receiver Operating Characteristic (ROC) analysis

For systems that produce soft classifications a whole range of pairs can be computed instead of just one sensitivity/specificity pair by changing the threshold at which an object is considered to be positive or negative. This procedure leads to the construction of a Receiver Operating Characteristic (ROC) curve (Fig. 1.5), which is regularly used in radiological research[63]. From the ROC curve the area under the curve (AUC) can be computed, which summarizes the overall performance of the system as the probability that, given a random pair of a positive and negative object, the positive object has a higher score. Note that differently shaped ROC curves can give the same AUC. In these situations a partial AUC can therefore be more informative, for example in screening programs where the risk of false-positives should be minimized.

In literature often ROC curves have not been computed directly from a set of scores with corresponding labels. Instead, scores were fitted first to a Gaussian distribution. Although fitting reduces noise and is based on statistical decision theory[64], it is not required for ROC construction[65] and can sometimes unrealistically change the shape of the curve. ROC curves in this thesis are all directly constructed from the scores.

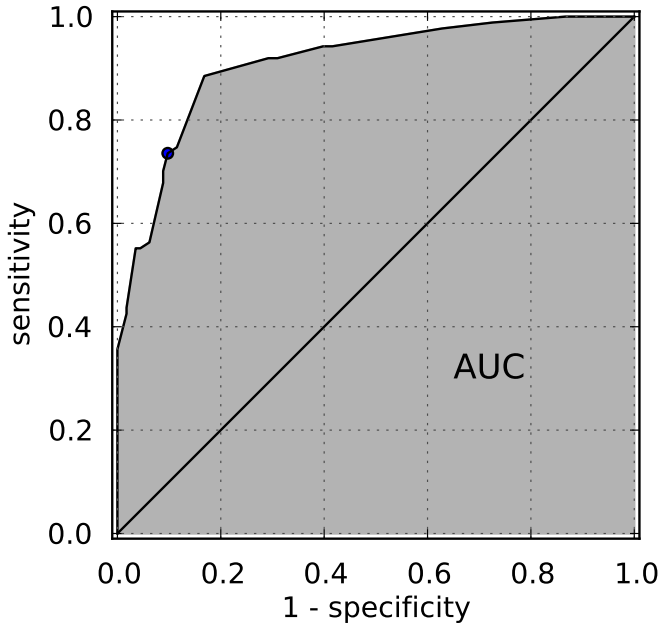


Figure 1.5: Example of Receiver Operating Characteristic (ROC) analysis. Performance of human readers against sputum culture reference (from Chapter 6). The shaded area is the Area under the ROC curve (AUC), a measure of overall system performance. The marker indicates the operating point of Table 1.2.(b)

Comparison with human readers

When developing an automatic method, it is not only important to compare it to other existing methods, but also to determine the performance with respect to human readers. In many tasks it is impossible for a computer algorithm to achieve a perfect score, because there is no consensus between humans in such a task. Especially when the reference standard is also set by a human observer, it is necessary to compare the performance of one or more other human observers to the automatic system. Another reason why it is important to compare the method to human readers is that their skill level can differ substantially. For example in a task where observers had to detect tumor masses in mammograms substantial differences between radiologists, but also between radiologists and non-radiologists, were found[66]. This implies that, although the CAD system does not achieve human expert performance, it can already replace less experienced readers.

Statistical testing

To conclusively show differences between two methods statistical significance testing is required, especially when differences are small. In this thesis most of

the comparisons between methods are performed on the same dataset. The advantage of this is that paired tests can be used, which analyze differences per case and therefore are in general more powerful than unpaired tests. Both Wilcoxon signed rank tests, and paired Student's t-tests were used in this thesis for assessing significance of differences between outputs of two methods.

In ROC analysis such tests cannot be used directly, because the main derived performance measure, the AUC, gives only one number per dataset. Although a specific test is available to compare AUC values, based on (parametric) fitting of the ROC curve, a nonparametric method was used throughout this thesis. Nonparametric methods do not assume a specific statistical distribution of the scores and thus can be in principle used on any set of outcomes. One of the most well-known methods is the bootstrap procedure[67]. In this method a set of bootstrap samples is created from which summary statistics such as confidence intervals can be computed. One bootstrapped sample consists of a set of objects sampled from the original set with replacement. Each bootstrap sample has the same size as the original set, and as such duplicates of original samples occur. In each bootstrapped sample the AUC (or other statistics) can now be computed. In ROC analysis each of the bootstrap samples provides an estimate of the AUC. By using exactly the same original samples in each bootstrapped sample to compute AUC values for two methods a paired test can be used to determine significance. The simplest of these paired tests is to compute a p-value directly by counting the percentage of samples in which one method performs better than the other[68].

Besides discarding the assumption of a Gaussian distribution on the scores, an important advantage of bootstrap analysis is that is applicable in complex experimental setups. A common example is the existence of correlation between samples, such as when multiple lesions or pixels of one image, are analyzed together. Case-based bootstrapping is then used to correct for correlation[69]. Dealing with such situations is difficult with standard tests, although for specific situations specialized tests exist, such as Multiple Reader Multiple Case (MRMC) analysis[70].

1.4 Automatic analysis of chest radiographs

The automatic analysis of CXRs has a long history and in fact started the field of computer-aided diagnosis[71]. Extensive reviews of the field are given in van Ginneken et al.[72] and Katsuragawa and Doi[73]. A recent overview of CAD research, that is concerned with other tasks than nodule detection, is given in van

Ginneken et al.[74]. Here we summarize the most important developments in the field and the work specifically related to TB detection.

Besides the detection of abnormalities, two topics have drawn the most attention: the segmentation of anatomical structures, and the problem of superimposed structures. The segmentation of the lung fields is required to limit subsequent analysis to this region only. Many of the early methods used unsupervised, rule-based methods, such as thresholding, region growing, edge detection, and shape fitting to segment the lungs[72]. More recent methods tried to classify each pixel in the image as lung or not lung, or applied statistical shape and appearance models to detect the lungs[75–77]. These latter methods have in general been more successful in providing accurate segmentations. In the normal, non-pathological lung, performances similar to humans have been achieved. The segmentation of pathological, possibly severely deformed, lungs has only received limited attention. Ribs are one of the most prominent structures in the CXRs and their segmentation has received considerable attention, in which a similar wide range of techniques as for the lung fields has been employed[72,78,79]. Segmentation of the clavicles has received less attention so far[76,80] and this topic is addressed in this thesis.

One of the main difficulties in analyzing the CXR is the superimposition of multiple structures at the same point in the image as a consequence of the 2-D projection. This property, designated *anatomical noise* by some authors[81], has been identified as one of the reasons for radiologists to miss abnormalities[82,83] and an increase of false positives in automatic methods[84]. Recently it has been shown that software that suppresses bony structures in the CXR can improve the radiologist's performance to detect nodules[85–87]. Several studies have addressed this problem. They can be divided in methods that use unsupervised filter techniques[88,89], supervised filter techniques[90,91], methods that avoid the bony structures altogether[92], methods that subtract the contralateral lung fields[84,93,94], and methods that subtract specific models of bony structures[80]. In Chapter 4 a method is presented that suppresses elongated structures, such as ribs and clavicles, in the CXR.

Concerning the detection of abnormalities in CXRs, nodules - possible lung cancers - have received most attention in CAD literature. Although this type of abnormalities is relatively uncommon in clinical practice[95], missing them has grave consequences both clinically (early detection is important for the prognosis) and legally (missed lung cancers are an important reason for litigation in

the United States[96]). Therefore many methods for detecting them have been developed and a number of commercially available software packages are on the market which can aid radiologists. An extensive overview of automatic methods that detect nodules is provided in de Boo et al.[97]. Analysis of diffuse (textural) abnormalities, which can occur in many lung diseases, has also been extensively investigated. Textural abnormalities refer to changes in appearance and structure of a region in the image compared to normal regions and can be measured by changes in the distributions of pixel densities in the region[53]. Texture analysis for the detection of abnormalities in CXR has a long tradition, going back to the 1970s[98]. Since then numerous techniques to measure texture changes have been developed and were applied to CXRs, including co-occurence matrices[99], power spectrum analysis[100], fractal analysis[101], multiscale Gaussian derivatives[102], and profile analysis[103]. An unusual shape of the lung fields can also be an indication of the presence of pathology and can be measured automatically, for example enlargement of the heart[104] or changes in the shape of the lung fields due to chronic obstructive pulmonary disease[105]. Both the detection of textural and nodular lesions are employed extensively in this thesis.

The specific detection of TB in CXRs has received less attention. A recent overview of automatic TB detection in CXRs is given in Jaeger et al.[106]. Several papers mentioned in the literature provide outlines of algorithms and automatic systems, but presented no quantitative evaluation of performance[107–109]. These papers are not mentioned here, because it is impossible to judge their value for automatic TB detection. Koeslag et al.[110] used template matching in the Fourier domain to determine the presence of miliary TB. In a set of 120 images they reported a high sensitivity of 94% with a moderate specificity of 68%. van Ginneken et al.[111] used texture analyis for TB detection. Lung fields were automatically segmented and divided in small regions. Texture features in these regions, based on moments of distributions of Gaussian derivative filtered images, were computed. The scores for classified regions were combined into one image score. In a set of 388 images an area under the ROC curve (AUC) of 0.82 was reported. Arzhaeva et al.[112] computed dissimilarities between distributions of features of an image to a set of prototype images. In a set of 471 CXRs an AUC of 0.83 was achieved. Lieberman et al.[113] used digital subtraction of follow-up CXRs and computed texture features on the subtraction images to monitor drug response. In a set of 200 images they found that the texture measures corresponded to visual improvement of the CXR. Shen et al.[114] detected and segmented cavities using mean shift

segmentation and active contour modeling. Several features of the candidate cavities were then computed to reduce false positives using a Bayes classifier. In a set of 149 images, 20 containing cavities, they reported a sensitivity of 82% with a false positive rate of 0.237/image. Tan et al.[115] analyzed the distributions of intensities in interactively segmented lung fields using the first three moments and the entropy. In a dataset of 95 images they found an area under the ROC curve (AUC) of 0.928. Jaeger et al. computed a number of different texture feature sets in automatically segmented lungs to detect TB[116]. In a dataset of 138 CXRs they found an AUC of 0.83 using SVM based classification. Note that directly comparing performances of these methods is not possible because different datasets from different populations were used in the studies.

1.5 Thesis outline

Several challenges remain to be addressed before an automatic system for TB detection on CXR can be used in practice. These challenges include basic image processing problems, such as anatomical noise and segmentation, the presence of image artifacts, the detection of specific types of abnormalities, and a large scale evaluation of the system. This thesis follows the general design of CAD systems as presented before. It concerns preprocessing, segmentation of anatomical structures, feature design, a description of the full CAD system, and an evaluation of the CAD system in a screening setting.

The outline of this thesis is as follows. In chapter 2 a method is described that detects and removes foreign objects, such as brassier clips and zippers, from CXRs. These objects complicate further automatic analysis and are regularly encountered in practical screening programs. Chapter 3 describes a method to segment the clavicles in CXRs. The clavicles are the most dominant structure in the upper lung region, the same region where TB is most often located. Segmentation was performed by a combination of several state-of-the-art methods. In chapter 4 a method based on unsupervised techniques was presented to suppress elongated structures, such as ribs and clavicles, in CXRs. The method was evaluated by determining its effect on the conspicuity of ribs, clavicles, and catheters. Chapter 5 describes the computation of a novel feature that measures asymmetry in the image. Deviations from expected symmetry can indicate the presence of pathology. The method was evaluated by detecting TB related pathology and its effects on the visibility of nodules. In chapter 6 a complete CAD system for TB detection is described that combines the outputs of several systems detecting different

kinds of abnormalities. The combination ensures that the system performs consistently across different screening populations. In chapter 7 the CAD system is applied to a large TB screening database, and its potential to triage (filter) active TB cases was assessed.

Foreign object detection and removal

2

Laurens Hogeweg, Clara. I. Sánchez, Jaime Melendez, Pragnya Maduskar, Alistair Story, Andrew Hayward, and Bram van Ginneken

Original title: Foreign object detection and removal to improve automated analysis of chest radiographs

Published in: Medical Physics 2013; 40(7):071901

Abstract

Chest radiographs commonly contain projections of foreign objects, such as buttons, brassier clips, jewelry or pacemakers and wires. The presence of these structures can substantially affect the output of computer analysis of these images. An automated method is presented to detect, segment and remove foreign objects from chest radiographs.

Detection is performed using supervised pixel classification with a kNN classifier, resulting in a probability estimate per pixel to belong to a projected foreign object. Segmentation is performed by grouping and post-processing pixels with a probability above a certain threshold. Next, the objects are replaced by texture inpainting.

The method is evaluated in experiments on 257 chest radiographs. The detection at pixel level is evaluated with ROC analysis on pixels within the unobscured lung fields and an A_z value of 0.949 is achieved. FROC analysis is performed at the object level, and 95.6% of objects are detected with on average 0.25 false positive detections per image. To investigate the effect of removing the detected objects through inpainting, a texture analysis system for tuberculosis detection is applied to images with and without pathology and with and without foreign object removal. Unprocessed, the texture analysis abnormality score of normal images with foreign objects is comparable to those with pathology. After removing foreign objects, the texture score of normal images with and without foreign objects is similar, while abnormal images, whether they contain foreign objects or not, achieve on average higher scores.

We conclude that removal of foreign objects from chest radiographs is feasible and beneficial for automated image analysis.

2.1 Introduction

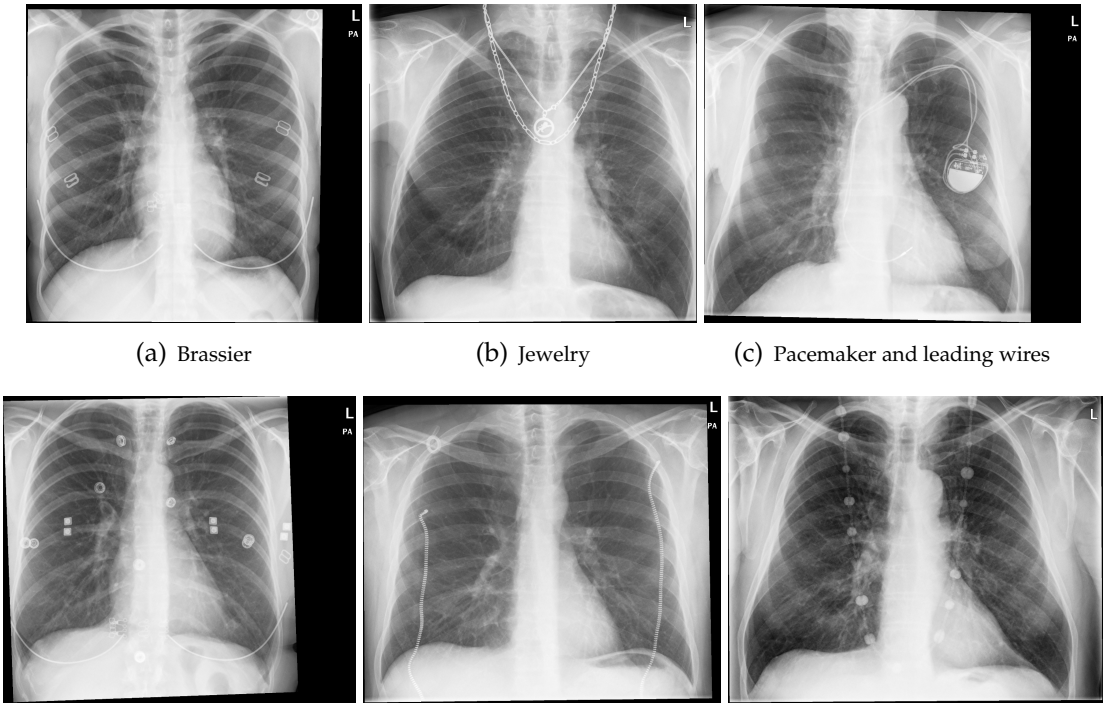
Algorithms for automated analysis of images often require a minimum quality of the input. Scientific studies that evaluate the performance of automated image analysis algorithms often exclude images of poor quality, either by removing them manually from the dataset or by using an automated algorithm to determine if the image quality is sufficient. Such quality assessment algorithms have been proposed for e.g. retinal images[117] and fingerprints[118].

In some situations detection and exclusion of low quality images may be sufficient, for example when it is possible to directly request a new acquisition. However, it will often be unfeasible or unacceptable to acquire a new image. An example is a screening program where participants might not be motivated to have a second exam taken. Bedside radiographs often contain foreign objects, like catheters, which cannot be removed before acquisition. It may also be the case that image quality is sufficient for reading by human experts, who can readily ignore and dismiss the artifacts, but automated analysis by computers will produce false alarms. If the locations of artifacts are known, one could attempt to deal with the issue by ignoring the computer output in areas affected by the artifact. A more ambitious strategy is to try to remove the artifacts. In this work we attempt to remove the artifacts in order to approximate the unaffected image as closely as possible.

Our focus is on detection and removal of foreign objects in chest radiographs. We use data from a large-scale screening program for tuberculosis. In a representative sample of 1,000 chest radiographs taken from this screening program almost 20% of the lung fields contained one or more foreign objects. Fig. 2.1 shows a number of different types of objects that were found to be present in these images. People are asked to remove their coats and any heavy chains, but are not required to fully disrobe; disrobing is time consuming and for some participants a barrier to participate.

To the best of our knowledge the topic of foreign object removal in chest radiographs has not been investigated before. The most closely related work in chest radiography image analysis has been done on automatic detection of catheter tips[119] (without the aim of actually removing the catheters) and the removal of anatomical structures such as ribs and clavicles by suppressing them[90,91].

More generally, the restoration of digital images has received considerable attention in the recent literature. Many older studies deal with reverting the effects



(a) Brassier (b) Jewelry (c) Pacemaker and leading wires

(d) Buttons and brassier (e) Zippers (f) Jewelry

Figure 2.1: Foreign objects in chest radiographs from Find and Treat database.

of noise such as blurring and speckle[120,121]. For the purpose of this paper we are especially interested in methods that are able to fill large holes (left by the foreign objects) in the image. Lee et al.[122] used bilinear interpolation along the short axis of elongated holes to remove hairs from dermoscopy images. One of the first automatic methods that uses texture synthesis to fill holes is the work of Efros and Leung[123] who employed nonparametric sampling. Bertalmio et al.[124] improved on this method by employing techniques derived from inpainting. Inpainting is a concept that originates from the restoration of paintings where the goal is to repair damaged parts in a visually convincing way[125]. Digital inpainting is described as the process where in a first step structural elements are continued into the holes, subsequently color is added to the still missing areas and finally texture is added. Bertalmio et al.[124] used anisotropic diffusion to propagate linear structures along isophotes. In an improved version texture was also added[126]. Criminisi et al.[127] further improved on this approach by removing the need to separate the image into its structural and texture component before inpainting. Both methods rely on nonparametric texture sampling to synthesize missing texture. Texture sampling can also be used as a standalone inpainting technique if

the restoration of structural elements is of less importance.

The work most similar to our paper, in its aim to improve automatic processing of medical images by repairing artifacts, has been done in the field of dermoscopy. In dermoscopic images linear structures such as hairs and rulers complicate (automated) analysis. Zhou et al.[128] used automatic line extraction followed by inpainting to remove the objects. Wighton et al.[129] used a similar approach but focused on the comparison between two methods to remove simulated hair. They found that using a specific type of inpainting, so called exemplar based inpainting, yielded images more similar to the original unaffected ones than using the linear interpolation algorithm DullRazor[122]. In a more recent comparison it was found that fast marching inpainting[130] performed better than linear interpolation, non-linear diffusion and exemplar-based inpainting in hair removal judged by segmentation performance and texture analysis[131]. In this work we employ a modified version of the texture synthesis method described by Efros and Leung[123] to inpaint areas where artifacts are superimposed on relevant image structure in chest radiographs.

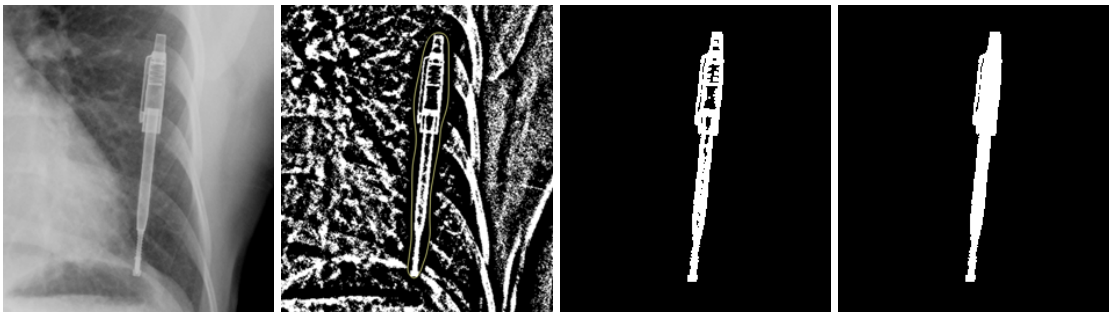
The paper is organized as follows. In Sect. 2.2 the data used in this study is described. Sect. 2.3 details the steps to automatically detect, segment and remove foreign objects from chest radiographs. The experiments to evaluate the method, in terms of detection and segmentation performance and in terms of its effect on subsequent use of the repaired images in a CAD system, are presented in Sect. 2.4 and result are given in Sect. 2.5. Sect. 2.6 discusses the results and Sect. 2.7 concludes.

2.2 Data

For evaluation of the detection and removal of foreign objects a large chest radiograph database from a tuberculosis screening program for high risk groups in London was used[5]. All images were digital and acquired with a single unit (DigitalDiagnost Trixel, Philips Healthcare, The Netherlands). In this work all images were scaled to a width of 1024 pixels, corresponding to a resolution ranging from 0.22-0.38 mm per pixel depending on the size of the original image. Although the original resolution is about a factor two higher (0.144 mm pixel size), we have found that this resolution is sufficient for the detection of tuberculosis related abnormalities in chest radiographs. This resolution is also sufficient for the detection and segmentation of foreign objects. If desired the inpainting could also be applied to full resolution images.

Set	Purpose	Content
A	Training set for foreign object detection	100 normal images with 331 foreign objects
B	Evaluation set for foreign object detection	107 normal images with 271 foreign objects
C	Training set for texture analysis	90 normal images, 48 with tuberculosis related pathology
D	Evaluation set for texture analysis	set B + 100 normal images without foreign objects + 50 with tuberculosis related pathology (8 containing foreign objects)

Table 2.1: Sets of images used for training and evaluating the algorithm.



(a) input image (b) local normalization (c) segmentation (d) hole filling

Figure 2.2: Semi-automatic segmentation of foreign objects in chest radiographs. (a) A part of a chest radiograph containing a foreign object. (b) Local normalization is applied to increase the contrast between the dense object and its direct surroundings. (c) The object is roughly outlined and the outlined region is thresholded and the correct connected components are selected. This segmentation is then manually post-processed with a painting tool. (d) When deemed necessary, interior pixels are added by hole filling.

Approximately 20% of the images in this database contain foreign objects. A subset of this database was used to evaluate the automatic detection of foreign objects and its effect on automatic detection of textural abnormalities. Table 2.1 lists four datasets defined for this study. Set A and B were used to train and evaluate the detection of foreign objects (Sect. 2.5.2), set C and D were used to train and evaluate the detection of textural abnormalities with and without foreign object removal (Sect. 2.5.3). Images in datasets A-D were randomly selected from the full database.

For training of the detection system precise annotations of foreign objects in the chest radiographs are required. These were obtained semi-automatically (see

Fig. 2.2). First, the contrast between the high density object with the background was further enhanced by local normalization[132] at a scale of $\sigma = 8$ pixels. This scale was determined by visual inspection and isolates high density objects from the background (Fig. 2.2(b)). The human operator set an appropriate threshold on this image and outlined the region of a foreign object. This region was masked and connected component analysis was used to quickly select all parts of the object to be segmented. The resulting binary mask was adjusted with a paint tool to obtain a precise segmentation and, if necessary, hole filling was applied. Objects with an area smaller than 120 pixels were excluded. This prevents detection of very small objects that are usually spurious. Using this procedure, 331 and 271 foreign objects were annotated in set A and B respectively, ranging in area from 120-6316 pixels.

2.3 Methods

The method starts with automatic segmentation of the unobscured lung fields. The proposed method consists of three stages (only performed inside the lung fields): detection of pixels belonging to foreign objects, segmentation of the objects, and removal of the object using inpainting. Both foreign object detection and lung field segmentation use pixel classification. An overview of the method is given in Fig. 2.3. The following paragraphs describe the method in detail.

2.3.1 Lung field segmentation

The detection and removal of foreign objects is limited to the unobscured lung fields as this is the main area of interest for many other applications of automated analysis. An automatic lung segmentation algorithm based on pixel classification was used to find the lung fields[76]. The system was trained with 500 training images where lung contours were manually outlined. These images were consecutively selected and some of them contained foreign objects and/or pathology. These images were not further used in the rest of the study. A number of small changes were made with respect to the system described by van Ginneken et al.[76]. Features were calculated at images of 1024 pixels wide (instead of 256 pixels), and resulting images are 512 pixels wide (instead of 256 pixels). As feature set we used Gaussian derivative and Hessian based features on the original images as described in Sect. 2.3.2. Pixels were classified using a k-Nearest Neighbor classifier ($k = 15$). The resulting probability map was slightly blurred ($\sigma = 0.7$ pixels), thresholded at a probability of 0.5, and morphologically closed with a spherical

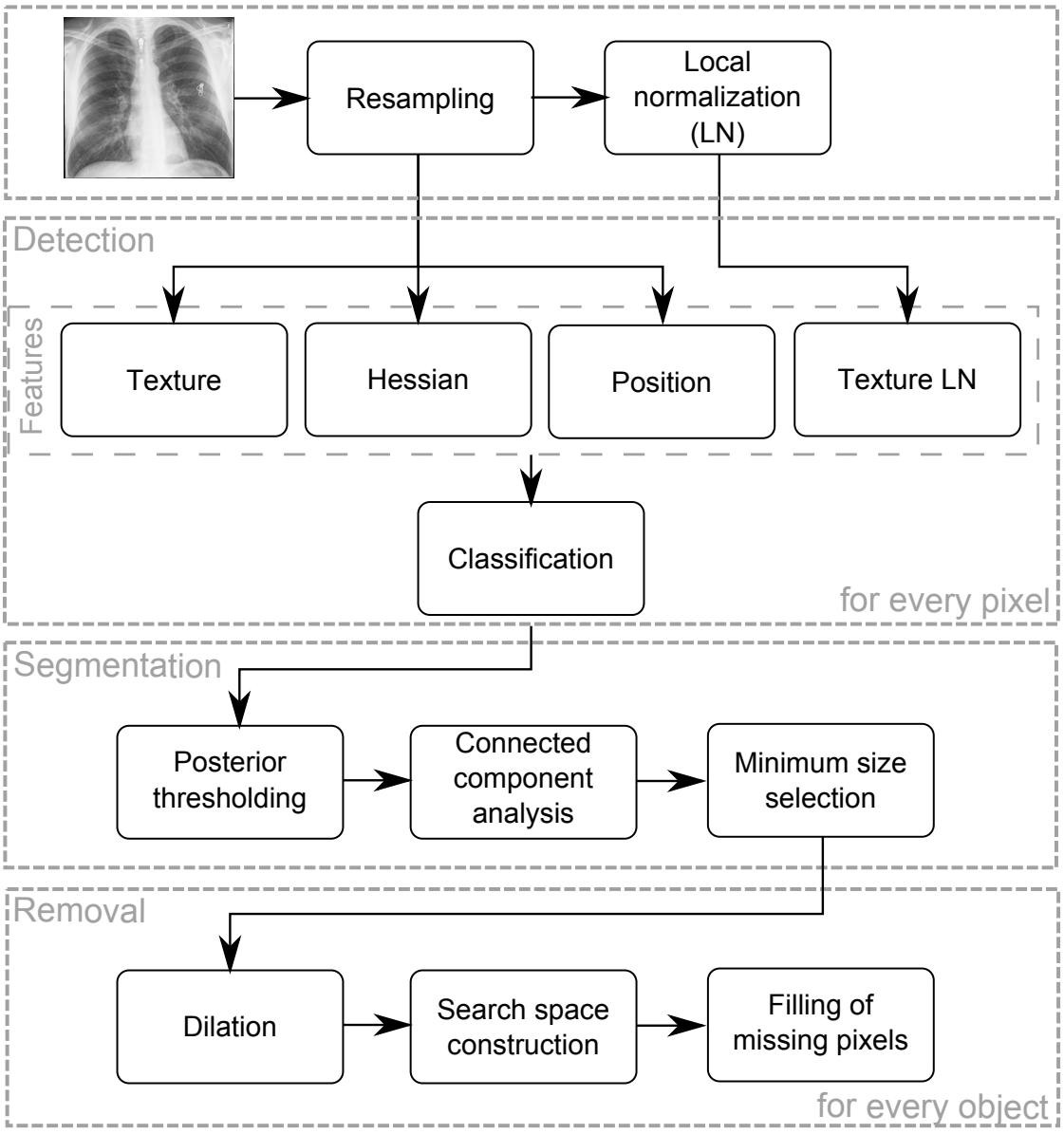


Figure 2.3: Overview of method. The method starts with resampling to a width of 1024 pixels and local normalization. Foreign objects are detected by computing features for every pixel inside the lung fields, followed by supervised classification. Objects are segmented by thresholding and connected component analysis. Removal of foreign objects is performed using a texture synthesis method which fills missing pixels with values obtained from the best matching patch in a search space around the object. See the text for details of each step.

kernel ($r = 10$ pixels). To segment the lungs the two largest components were selected.

2.3.2 Detection of foreign objects

Pixel classification (PC)

In this methodology the segmentation problem is recast into a pattern classification task[51,54]. A number of continuous characteristics (features) are calculated for a number of samples (positions, pixels) in an image. A classifier is trained using labeled samples from a database of training images. Examples of labels are inside/outside foreign objects and inside/outside the unobscured lung fields. The classifier provides the mapping from features to class labels. In this work we use classifiers that provide a posterior probability that indicates the likelihood for a sample to receive a label. Test images can be segmented afterwards by computing the features for each position and applying the classifier.

Features

Three types of features were calculated for each sample: texture features based on Gaussian derivatives (on original and locally normalized images), features derived from the Hessian matrix and position features. First each image is resized to a width of 1024 pixels. To capture local image structure[133] the output of Gaussian derivative filtered images of order 0, 1, 2 (L, L_x, L_y, L_{xx}, L_{xy}, L_{yy}), at scales 1, 2, 4, 8, 16, 32 and 64 pixels were calculated. The small scales provide information about the fine image structure and the larger scales about the neighborhood of the pixel. For speed improvement the recursive implementation described by Deriche[134] was used. Foreign objects typically have a high density compared to their background. As explained in Sect. 2.2 applying local normalization to the image will improve the contrast of the object with the background. To reflect this in the feature set, images were locally normalized with $\sigma_{LN} = 8$ pixels, the same scale as used for annotation, and features derived from the output of Gaussian derivative filtered images of order 0 and 1 (L, L_x, L_y) at scales 1 and 2 were added. The use of small scales specifically enhances strong edges, which are characteristic of foreign objects encountered in our dataset. Certain types of foreign objects consist of thin, elongated lines. Hessian matrix derived features were used to detect the presence of these line like structures[135]. Considering the two eigenvalues of the Hessian matrix λ_1, λ_2 with $|\lambda_1| > |\lambda_2|$, two measures were derived: (1) $\sqrt{(\lambda_1^2 - \lambda_2^2)}$ to extract the lineness of the local image structure and the largest

absolute eigenvalue $|\lambda_1|$ to indicate the strength of the response. These Hessian features were calculated at scales 1, 2, 4, 8 and 16 pixels. Finally the x and y position were added to account for the difference in background appearance across the lung. In total 61 features were computed per sample.

Classification

The training set was constructed by sampling inside the lung fields 100% of the available positive (foreign object) pixels and a random 0.5% of the available negative pixels per image. Only a small percentage of normal pixels was sampled because they occur much more frequently in the original image. To reduce computation time test images were subsampled by a factor 2 resulting in probability maps with a width of 512 pixels. For segmentation, removal, and evaluation the probability maps were resized back to a width of 1024 pixels.

All features were normalized to zero mean and unit standard deviation before classification. A number of different classifiers for foreign object detection were tested. In the following description of the classifiers we use boldface to denote vectors or matrices, $\mathbf{x}$ for feature vectors (samples), and $y \in \{-1, +1\}$ for sample labels of classes 1 and 2. All classifiers output posterior probabilities.

Linear Discriminant Analysis Fisher's linear discriminant analyisis (LDA) is a discriminative method which is commonly used in CAD applications for its speed and simplicity. The method finds a linear discriminant function in a supervised fashion assuming that the class density $f_c(\mathbf{x})$ for each class c is a multivariate Gaussian distribution with the same covariance matrix Σ:

$$f_c(\mathbf{x}) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{m}_c)^T \Sigma^{-1}(\mathbf{x} - \mathbf{m}_c)) \tag{2.1}$$

with $c = 1, 2$ and $\mathbf{m}_c$ the class mean. The posterior probability for $y = 1$ given $\mathbf{x}$ is given by

$$P(y = 1|\mathbf{x}) = \frac{\pi_1 f_1(\mathbf{x})}{\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})} \tag{2.2}$$

where π_c is the prior probability of class c.

Nearest mean classification Nearest mean classification assigns to a test point the label of the nearest class mean. The posterior probability is calculated in a similar way as LDA, using Eq. 2.2, under the assumptions of multivariate Gaus-

1. Given datapoint and label pairs $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)$ where $\mathbf{x}_i \in X, y_i \in Y = \{-1, +1\}$
2. Start with $H(\mathbf{x}_i) = 0$ and weights $w_i = 1/N$, $i = 1, ..., N$
3. Repeat for $m = 1, ..., M$
   (a) Find the optimal weak classfier h_m for (X, Y) and current weights w_i
   (b) Update strong classifier $H(\mathbf{x}) \leftarrow H(\mathbf{x}) + h_m(\mathbf{x})$
   (c) Update weights for examples $w_i \leftarrow w_i e^{-y_i h_m(\mathbf{x}_i)}$ for $i = 1, ..., N$

Figure 2.4: GentleBoost algorithm

sian class distributions with the same covariance matrix Σ and $\Sigma = \sigma^2 I$ where I is the identity matrix. Like LDA, the Nearest mean method is unable to model complex decision boundaries but it is fast, performs well in some settings, and makes no assumptions on the data.

k-Nearest Neighbor k-Nearest Neighbor (kNN) classification is a nonparametric method where the decision boundary is constructed locally in an area around the query sample. The posterior probability for $y = 1$ given $\mathbf{x}$ is given by

$$P(y = 1|x) = \frac{k_1}{k} \tag{2.3}$$

where k_1 is the number of samples among the k nearest neighbors with label $y = 1$. The Euclidean distance was used as a distance measure in this work. kNN has the attractive property that with increasing training size the conditional error approaches the Bayes error[54]. To speed up the classification the tree-based implementation by Arya et al.[136] was used. This implementation uses an approximate solution controlled by the variable ϵ, which ensures that the approximate nearest neighbors are no more than (1+ϵ) times the distance away from the query point than the actual nearest neighbors. ϵ was set to 2 in this work.

GentleBoost The GentleBoost algorithm belongs to the family of ensemble classifiers and was described by Friedman et al.[58]. A number of weak classifiers h_m are sequentially combined into a strong classifier H using the algorithm shown in Fig. 2.4. The algorithm uses adaptive Newton steps in each round m to minimize

the weighted squared error

$$J_m = \sum_{i=1}^{N} w_i (y_i - h_m(\mathbf{x}_i))^2, \tag{2.4}$$

where $w_i = e^{-y_i H(\mathbf{x}_i)}$ are the weights, h_m the weak classifier, and N the number of training samples. The optimal weak classifier is then then added to the strong classifier H and the weights are adapted. For the weak classifier any suitable algorithm can be chosen, but for GentleBoost often a simple algorithm, such as regression stumps, is used. Regressions stumps are basically decision trees with one node and are defined as $h_m(\mathbf{x}_i) = a\delta[\mathbf{x}_i^f > \theta] + b\delta[\mathbf{x}_i^f < \theta]$, where f is the the feature number and δ the indicator function. The stump is optimized by finding the parameters $\{a, b, f, \theta\}$ that minimize J_m. A closed form solution for a and b can be derived and $\{f, \theta\}$ are found by exhaustive search[137]. The trained strong classifier H gives the log-odds of being in class y where $H(\mathbf{x}) = \log P(y|\mathbf{x})/P(-y|\mathbf{x}), y \in \{-1, +1\}$. The posterior probability for $y = 1$ given $\mathbf{x}$ is estimated using a sigmoid function as follows:

$$P(y = 1|\mathbf{x}) \approx \frac{1}{1 + e^{-H(\mathbf{x})}}. \tag{2.5}$$

GentleBoost has been shown to have improved performance compared to other classifiers such as AdaBoost[138]. For boosting algorithms in general, similar performance was found as for neural networks but with decreased training times[139].

Support Vector Machine The Support Vector Machine (SVM) constructs a hyperplane in a high-dimensional space in such a way that the distance between the hyperplane and the two classes is maximal, which makes it a maximum-margin classifier. The hyperplane is found by solving the following minimization problem[56]

$$\begin{aligned} \underset{\mathbf{w},b,\xi}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{2.6}$$

where $\mathbf{w}$ and b are the weights and the bias of the hyperplane, respectively. The parameter $C > 0$ controls the misclassification error induced by the slack variables ξ, which are introduced to allow for solutions when a hyperplane splitting

the classes does not exist. The function ϕ maps the feature vectors into a higher dimensional space where a hyperplance may be easier found. ϕ is related to a kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$. Different kernel functions can be used. Setting $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ gives a linear kernel. The Gaussian radial basis function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2), \gamma > 0$ is often used as a nonlinear kernel and introduces an extra parameter γ. SVM does not provide posterior probabilities $P(y = 1|\mathbf{x})$ directly, but these can be estimated from the distances of samples to the hyperplane by crossvalidation on the training set, for details see Chang and Lin[140]. Training times for SVM can be considerable as C and γ have to be determined in an exhaustive search procedure. Testing times are typically much faster and related to the number of support vectors needed to define the hyperplane.

2.3.3 Segmentation

The output of the pixel classifier is a probability for each location in the lung fields to belong to a foreign object. A binary segmentation is obtained by thresholding the probabilities of the pixel classifier using a threshold p_t. Subsequently connected component analysis (using 8-connectedness) is performed. Only objects with an area > 120 pixels were retained. The effect of p_t on the detection performance is evaluated in Sect. 2.5.2.

2.3.4 Removal of foreign objects

After detection, foreign objects were removed from the image to restore the appearance of the background lung. For removal we adopted a texture synthesis algorithm, which uses non-parametric sampling to recover the texture at the location of the removed foreign object. The removal is performed on images of 1024 pixels wide. Incorrectly segmented foreign object pixels at the boundary (false negatives), which typically have a high density, could be incorrectly used as example pixels for the removal algorithm and lead to artifacts in the restored image. To reduce this risk the detected objects are slightly dilated with 4 pixels before texture synthesis is performed.

For texture synthesis the method of Efros and Leung[123] was taken as the basis. Missing pixels are filled one at a time by finding pixels with similar neighborhoods and copying the value of the pixel with the best matching neighborhood. The neighborhood is defined as a square patch surrounding the pixel. We used patches of 11×11 pixels. Similar neighborhoods are determined by calculating distances between the patch for the missing pixel and all other possible neighbor-

hoods. The distance d is defined as the sum of squared differences (SSD) between the two patches. The SSD is only calculated using pixels that are known in both patches. To prevent the method from getting stuck in local minima a random patch is selected from a set of best matching patches with $d < (1 + e)d_{best}$, where d_{best} is the distance of the best matching patch and e is a threshold.

While the method is intuitively simple, it is in practice very slow because the whole image has to be searched for every missing pixel. Three modifications were made to make the method useful in practice. To speed up the search process of similar patches we used Approximate Nearest Neighbor (ANN) search[136], a fast version of k-nearest-neighbor (kNN) search (inspired by the use of Tree Structured Vector Quantization in Wei and Levoy[141]). The use of kNN search also replaces the parameter e from the original algorithm, which gives a variable number of best matching patches, by a fixed number of patches controlled by k. We used $k = 10$. In a structured image such as a chest radiograph similar patches are expected to be found close to the missing pixel, therefore we limited the search area to an area of 50 pixels around each of the objects to be filled. The considerable reduction of the size of the search space also improves computation times. Contrary to the original algorithm, which updates the search space after each iteration, the search space is only constructed once per object. Finally we took the output of the DullRazor[122] method as a pre-processed input image for texture synthesis. DullRazor performs bilinear interpolation at each missing pixel between the endpoints of the shortest ray crossing the hole. Eight rays were cast in equally spaced directions to determine the shortest ray. Prefilling the hole is necessary as kNN requires complete patches (with no missing pixels) to search with, while missing pixels typically also have missing pixels in their neighborhood.

2.4 Experiments

2.4.1 Foreign object detection: pixel based evaluation

The training set consisted of 59,887 pixels (49,268 pixels from background, 10,619 pixels from foreign objects) sampled from set A. The test set consisted of 108,052 pixels (47,743 pixels from background, 60,309 pixels from foreign objects) sampled from set B.

For detection of foreign objects we tested LDA, kNN classification, Nearest mean classification, a Support Vector Machine (SVM), and GentleBoost. Optimal

parameters and features were determined for the classifiers. From pilot experiments it was determined that a high pixel level specificity is required to limit the number of false positive object detections. Therefore the optimization criterion was set to the sensitivity at 0.995 specificity. The value of k in kNN was optimized in crossvalidation on the training set. Odd values of k in the range 1-301 were tested, and $k = 19$ was determined to be the optimal value. For SVM a Gaussian kernel function was used, the hyperparameters C and γ were optimized in a grid search by crossvalidation on the training set, according to the recommendations in Hsu et al.[142]. The optimal values of C and γ were 2^{-1} and 2^{-5}, respectively. GentleBoost used regression stumps as the weak classifier. A number of 1000 stumps were added to train the classifier. Feature selection was performed using Sequential Forward Selection (SFS)[143] for each of the classifiers. Feature selection for SVM was not performed, because optimal values for the hyperparameters would have to be determined for every different subset of features, leading to prohibitively long computation times.

Receiver operating characteristic (ROC) analysis was performed on the test set. The pixel based classifier performance was measured using the Area under the ROC curve A_z. Differences between A_z values were determined with bootstrapping[67], using 1000 bootstrap samples and a significance level $\alpha = 0.05$.

2.4.2 Foreign object detection: object based evaluation

Object detection performance is determined using Free Response Operator Characteristic (FROC) analysis[144]. For FROC construction criteria are needed for true positive (TP), false positive (FP) and false negative (FN) objects. The criteria are based on the fraction of positive pixels detected $\Omega = TP/(TP + FN)$. An object in the segmentation is TP when it overlaps with a foreign object of which at least 50% of the pixels are detected ($\Omega > 0.50$). If an object overlaps but less than 50% of the pixels are detected ($0 < \Omega < 0.50$) it is considered a FN. Objects in the reference standard which do not overlap with any object in the segmentation are also FN. Finally, an object in the segmentation is considered FP when $\Omega = 0$.

During construction of the FROC curve, p_t is lowered to obtain segmentations at different levels and determine the number of TP, FN and FP objects. The area of the segmented objects (connected components) typically grows when the threshold p_t is lowered, and may lead to merging of two or more FP objects into one and a counterintuitive reduction of number of FPs at a lower value of p_t. This has been noted as an issue in FROC analysis in several studies[145,146]. To prevent

this effect a local maxima detection scheme is used. FROC construction starts at $p_t = 1.0$. At each level of p_t FP objects are identified in the segmentation. For each FP object it is determined whether it overlaps with an object detected at a higher p_t. If it was previously detected it is ignored, otherwise the object is recorded as FP and assigned a score equal to the current p_t. In a similar way the TP and FN objects are assigned a score based on the level of p_t where they first appeared. In the reported results the step size for p_t was 0.05.

The object score is used to construct the FROC curve. Pilot results showed that many false positive responses of PC occur at the boundary of the lung where the high gradient transition between bone and lung tissue mimics that of foreign object and lung tissue. To reduce the effect of lung segmentation inaccuracies on the results, the region within a distance of 4 pixels (0.88-1.52 mm) from the automatically detected lung boundary was excluded from analysis.

2.4.3 Effect on textural abnormality detection

One of the goals of foreign object removal is to improve automated analysis of chest radiographs. Textural (parenchymal) abnormalities are a common disease pattern in chest radiographs and can be found in different types of diseases, such as pneumonia or tuberculosis. To evaluate the effectiveness of the foreign object detection and removal algorithm, a previously described CAD system for tuberculosis detection on chest radiographs was used[147,148]. Classification is based on texture analysis of small circular image patches placed inside automatically detected lung fields[76]. Features based on moments from Gaussian derivative filtered images are calculated for each patch and assigned a probability of abnormality using LDA. One texture score, reflecting the total load of abnormal patches in the image, was determined by integrating the individual patch probability scores. This integration was performed by calculating the 95th percentile of the cumulative histogram of patch probabilities[149] which provides a robust estimate of the relative area of lung affected by pathology.

The textural abnormality system was trained with images in set C. Textural abnormalities in this set were outlined to provide examples of abnormal patches. Examples of normal patches were only sampled from normal images. The total number of normal and abnormal patches used to train the LDA classifier was 122,932 and 4,958, respectively. The system was then applied to test images in set D.

The CAD system was applied two times; first using features derived from

#	Individual performance	Sensitivity at 0.995 specificity	Selection order in SFS	Sensitivity at 0.995 specificity
1	λ_1 (σ = 1 pixels)	0.356	λ_1 (σ = 1 pixels)	0.356
2	L_{yy} (σ = 2 pixels)	0.316	L_{xx} (σ = 4 pixels)	0.516
3	L_{yy} (σ = 4 pixels)	0.288	L_{yy} (σ = 4 pixels)	0.643
4	λ_1 (σ = 2 pixels)	0.285	L_y (σ = 64 pixels)	0.674
5	L_{yy} (σ = 1 pixels)	0.271	L_y (σ = 1 pixels)	0.743

Table 2.2: Analysis of feature performance on the pixel level with kNN as classifier. On the left are shown the five features which perform best individually. On the right the five features first selected using Sequential Forward Selection (SFS), in other words the best performing subset, are shown.

original test images and then from processed test images where foreign objects had been removed. Images in set D were divided into 3 groups: normal images containing foreign objects, abnormal images, and normal images without foreign objects. Texture scores were calculated for the 6 situations (original and processed images in the 3 groups) and displayed using box plots. Differences between means of the groups were compared using unpaired Student t-tests ($\alpha = 0.05$).

2.5 Results

2.5.1 Foreign object detection: pixel based evaluation

All classifiers achieve a high classification performance. Differences are small, but significant differences in A_z were found between LDA, the best performing classifier, and the other classifiers. At high specificities (0.99-1.0), the operating region for segmentation, kNN and SVM are the best classifiers (Fig. 2.5). Nearest mean classification has the lowest performance but still reaches a high overall A_z value. Feature selection did not lead to improved sensitivity at the level of 0.995 specificity, except for Nearest Mean classification, which remained the worst performing classifier. Despite the lack of performance improvement, feature selection provides insight into how individual features perform. The five features performing best individually and the best subset of five features with kNN as classifier are shown in Table 2. The features that are most often selected, namely λ_1 and second order derivatives at small scales, reflect the presence of fine line-like structures. This is due to the fact that foreign objects are mainly thin elongated structures with sharp borders, specially highlighted at small scales. Features at larger scales also contribute to classification but mostly in combination with other features.

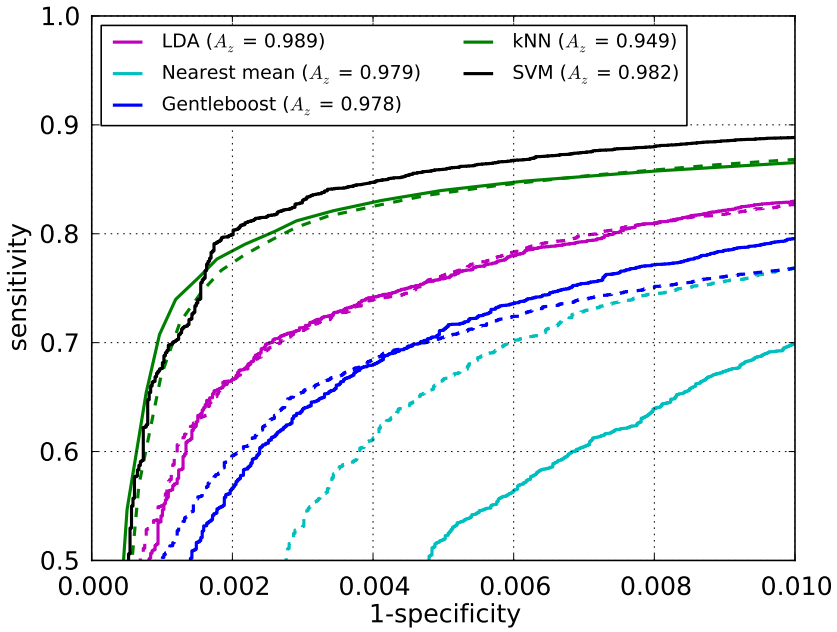


Figure 2.5: ROC Analysis of pixel classification. Shown are results for Linear Discriminant Analysis (LDA), Nearest mean, k-nearest-neighbor (kNN), GentleBoost, and SVM. The results with feature selection are indicated with dashed curves.. The ROC curve is shown at high specificities where the pixel classifiers operate.

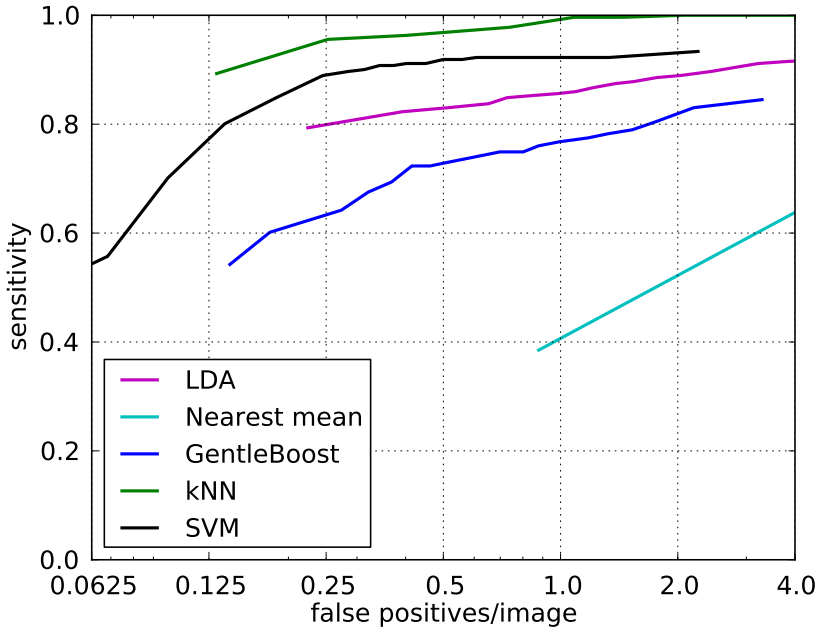


Figure 2.6: Performance analysis at the object level. FROC curve for different classifiers (see Fig. 2.5) in the 0.0625-4.0 FP/image operating range. The x-axis is logarithmically scaled.

2.5.2 Foreign object detection: object based evaluation

Fig. 2.6 shows FROC curves of the tested classifiers indicating object detection performance for the 0.0625-4.0 FP/image operating range. The left most point of a curve indicates the detection performance for the lowest probability map threshold ($p_t = 0.05$).

kNN showed superior sensitivities compared to the other classifiers in the analyzed FP/image range. SVM was the second best performing classifier. LDA showed worse performance than kNN and SVM. GentleBoost has reduced sensitivities compared to kNN, LDA, and SVM; while Nearest mean classification has markedly reduced performance compared to all classifiers. For the remaining experiments we used kNN. At a level of 0.25 FP/image 95.6% of the objects were successfully detected using kNN. This level corresponds to $p_t = 0.90$ and was used in subsequent experiments unless indicated otherwise. At $p_t = 0.90$ the corresponding per pixel sensitivity and specificity are 0.65 and 0.999, respectively. This operating point corresponds to the operating range where kNN performs best on the per pixel level (Fig. 2.5). Increasing the sensitivity at the object level of kNN beyond 95% finds only slightly more foreign objects at the expense of a large increase in the number of FP objects.

2.5.3 Effect on textural abnormality detection

Figs. 2.7 through 2.9 show an overview of the results of the detection, removal and effect on texture analysis of 12 selected cases. Fig. 2.7 and Fig. 2.8 show cases containing foreign objects, Fig. 2.9 cases without foreign objects. The fourth example in Fig. 2.7 displays an abnormal case containing extensive pathology. From top to bottom the original images are shown, followed by the manual segmentations of the foreign objects (the reference standard). The next row shows detection using pixel classification. The fourth row shows the segmentation as produced by thresholding the output of the pixel classifier with $p_t = 0.90$, connected component analysis, and retaining components with a minimum area of 120 pixels.

The fifth row shows the images with the segmented objects inpainted by texture synthesis. The last two rows show the output of the textural abnormality detection system on the original images, and on the processed images in which the foreign objects have been inpainted. Applying texture analysis to the original images leads to false positive responses due to the presence of foreign objects. The strong responses lead to texture scores in normal images that are similar to

those in abnormal images. After removing the foreign object by inpainting, the false positive responses have mostly disappeared, leading to markedly reduced texture scores. In cases containing no foreign objects the texture response is identical before and after removal, except for an occasional false positive response on the lung border.

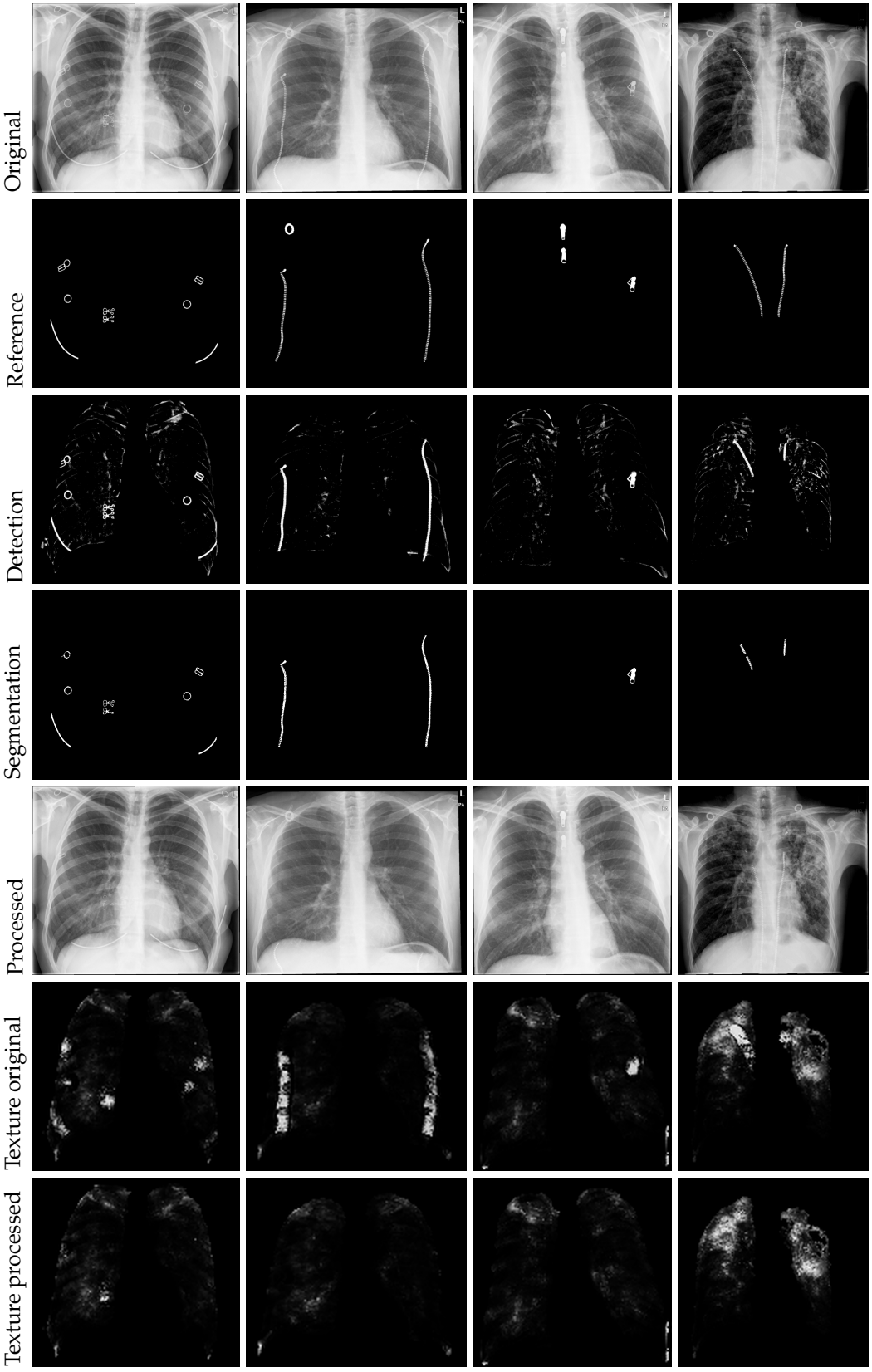


Figure 2.7: Illustration of the output of the algorithm for three selected normal cases and one abnormal case (4th column) with foreign objects. See text for explanation.

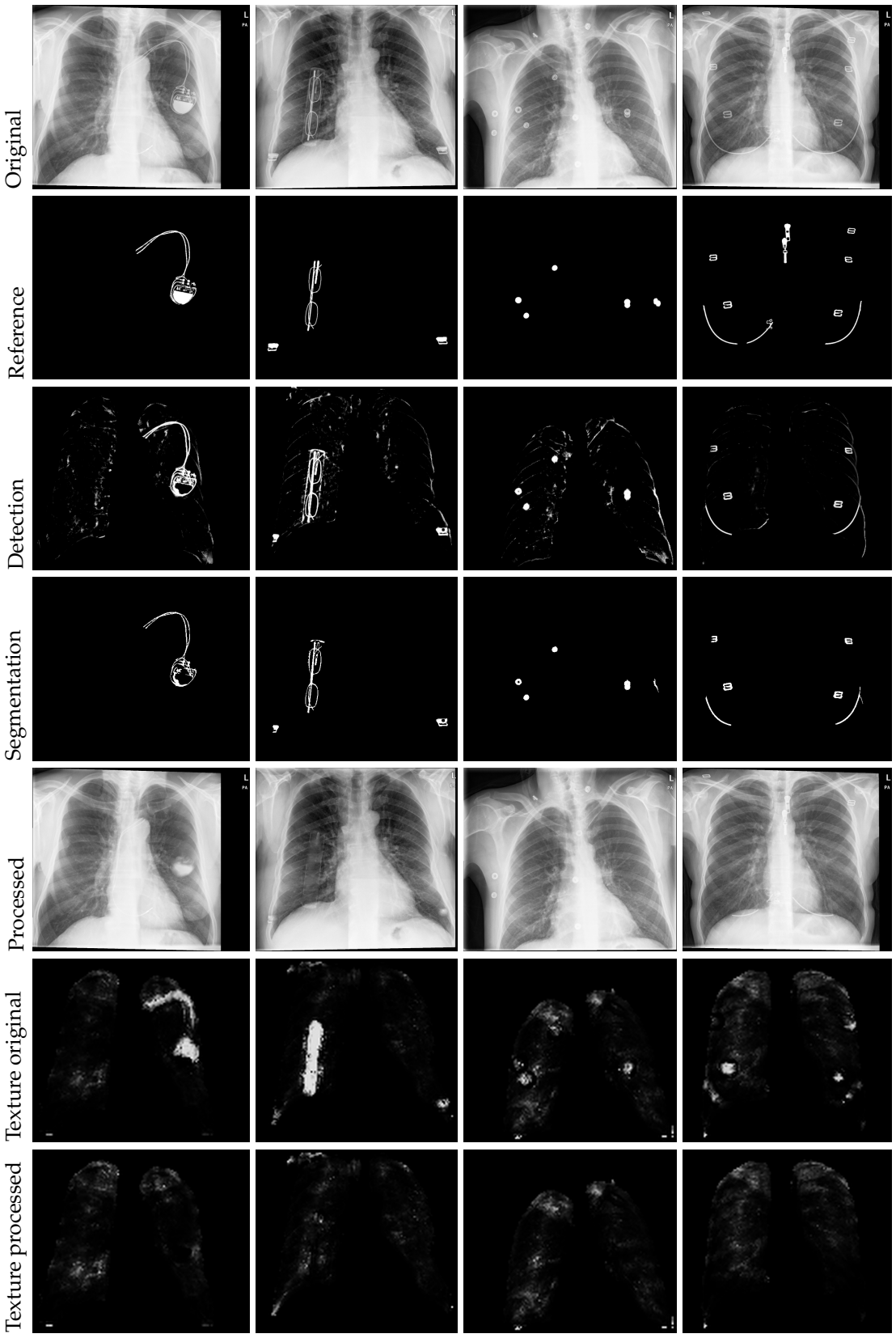


Figure 2.8: Illustration of the output of the algorithm for four selected normal cases with foreign objects. See text for explanation.

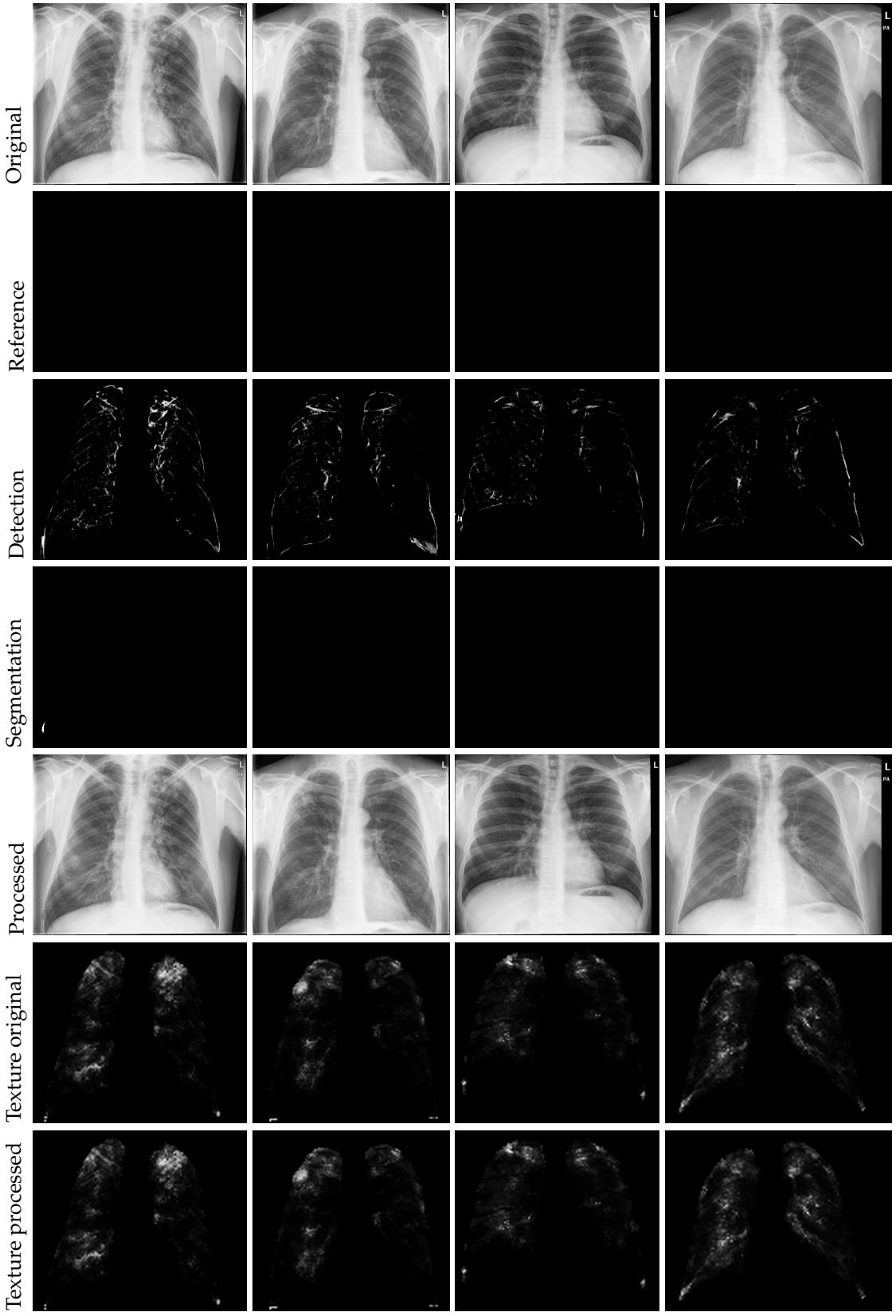


Figure 2.9: Illustration of the output of the algorithm for four cases with no foreign objects. The first two cases contain pathology (1st case: abnormality in upper left lobe, 2nd case: abnormality in upper right lobe just below the clavicle), the last two cases contain no pathology.

Close-ups of a number of foreign objects are shown in Fig. 2.10.

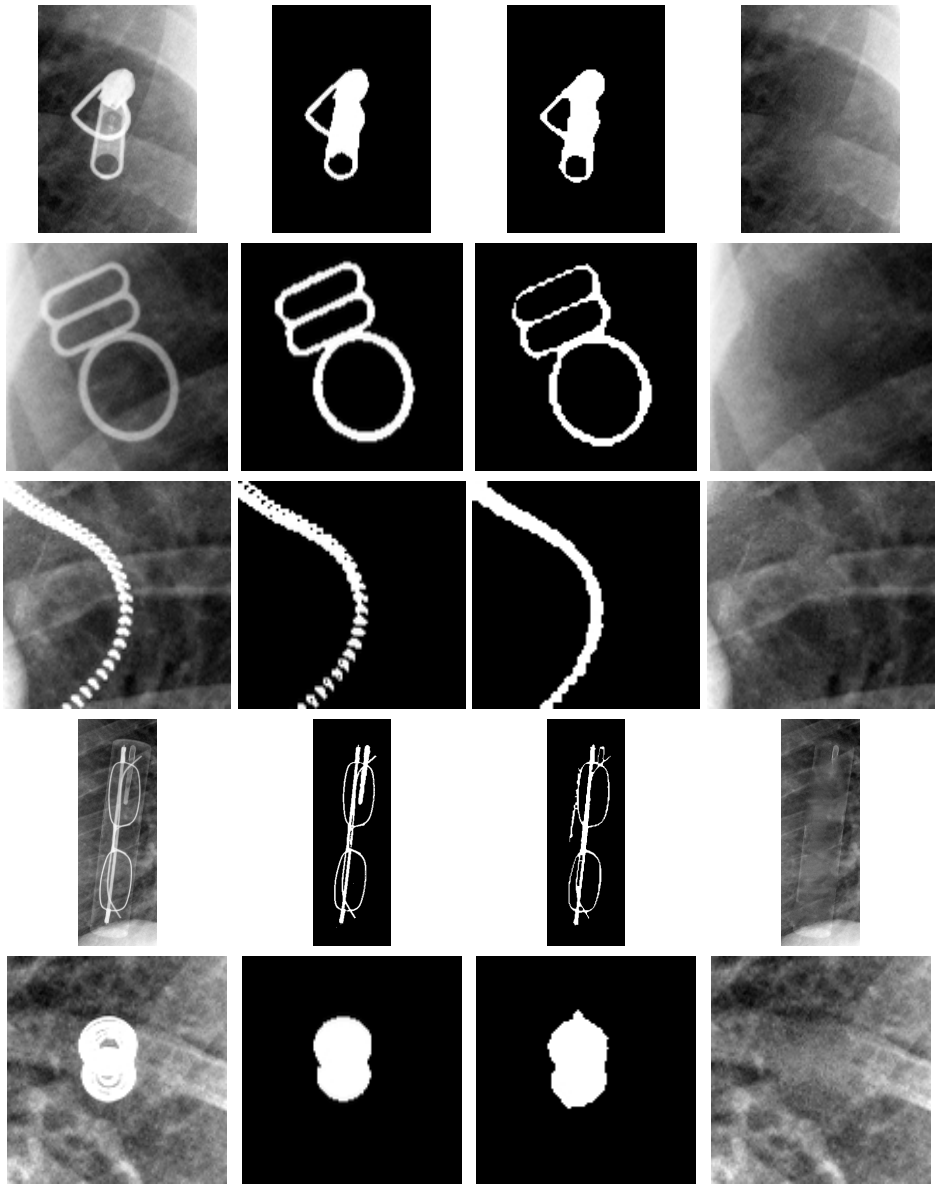


Figure 2.10: Close-ups of several foreign objects. The columns show respectively the original image, the reference annotation, automatic segmentation and result of removal using texture synthesis. From top to bottom a zipper slider, brassier clip, zipper, glasses with case, and a button are shown.

To quantify the effect of foreign object removal, CAD texture scores were compared between normal and abnormal images and before and after image restoration in set D. Boxplots[150] in Fig. 2.11 show that texture scores in normal images with foreign objects are higher than in normal images without foreign objects be-

fore processing (boxplot 1 and 5; $p < 0.05$). This reduces the ability of the textural abnormality system to discriminate between abnormal images and normal images with foreign objects as they show no significant difference in scores (boxplot 1 and 3; $p = 0.35$). After processing, the scores of normal images with foreign objects are reduced to a similar levels as normal images without foreign objects (boxplot 2 and 6; $p = 0.93$). The restoration has no noticeable effect on texture scores of images containing no foreign objects (boxplot 5 and 6; $p = 0.89$). A slight, but not significant, reduction is observed in the scores of abnormal images as some of them contain foreign objects (boxplot 3 and 4; $p = 0.63$). Further analysis of the abnormal cases showed a similar pattern as in the normal cases. After processing, the average score of the 8 abnormal cases containing foreign objects scores was reduced to a similar value as the average score of the 42 abnormal cases without foreign objects before processing ($p = 0.82$). This indicates that, although scores in abnormal cases are reduced, it can be predominantly ascribed to the foreign objects being removed, not to the removal of pathology. An example of an abnormal case with a zipper removed can be seen in Fig. 2.7, 4th column.

2.6 Discussion

A method was presented to automatically detect and remove foreign objects in chest radiographs and was evaluated on images acquired from a tuberculosis screening program. In screening practice the occurrence of foreign objects is not uncommon and automatic removal is a necessary prerequisite for other automatic processing of chest radiographs. The contribution of this paper is two-fold: 1) the application of state-of-the-art techniques for segmentation and image restoration to an unexplored application; and 2) the modification of an existing texture synthesis method which reduces computation time greatly and renders its use more practicable. In this section, the detection and segmentation performance in relation to the false negatives and false positives are discussed first. Then a discussion of the removal of the objects and its usefulness in practice follows. Finally the use of the system in a practical context is discussed.

Detection and segmentation of the foreign objects was in general quite accurate. Clearly delineated objects, such as metal objects (e.g. brassieres, coins, keys) were typically detected, accurately segmented, and removed by the algorithm. This is illustrated by the first 3 close-ups of foreign objects in Fig. 2.10 where after inpainting the previously affected area is largely artifact free. The last close-up of Fig. 2.10 indicates that a perfect segmentation is not required to obtain a convinc-

ing inpainting result. A slight oversegmentation of objects is not a problem as the FP pixels will be restored by the inpainting algorithm. To make the algorithm less sensitive to undersegmentation obtained segmentations were slightly dilated before inpainting. A larger dilation might further reduce the risk of FN pixels, but it will also make it difficult for the inpainting to succeed as larger holes lead to information loss about the local appearance.

The algorithm missed 12/271 (4.4%) of the foreign objects at the cut-off used for segmentation. The majority of them were small (10 FN objects with area < 400 pixels, 20-50 mm^2 depending on the original image size) and we expect their influence on the subsequent processing algorithms to be minor. Some categories of objects were less well handled. A number of large objects having uniform areas of density, such as pacemakers (Fig. 2.8, 1st case), were not completely segmented

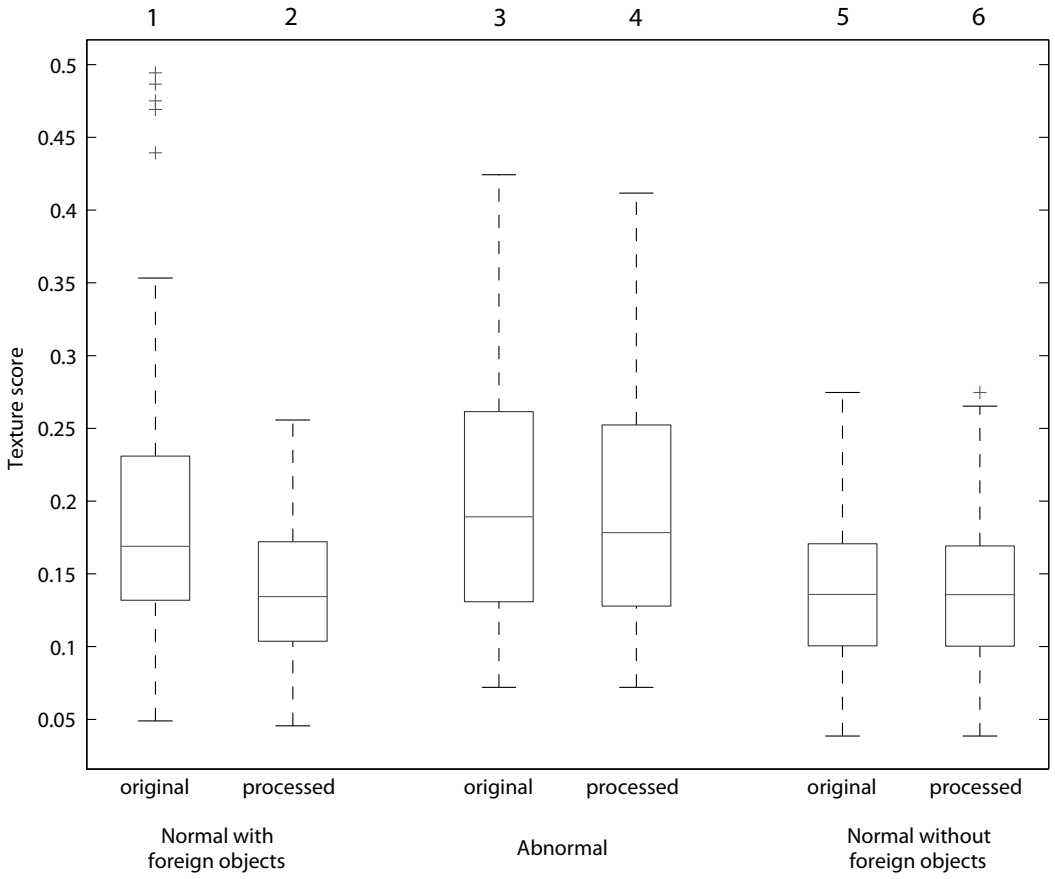


Figure 2.11: Effect of restoration on texture scores for normal images containing foreign objects, normal images without foreign object and abnormal images. Boxplots of the texture score for the three groups of original and processed images.

and removed. As there were only a limited number of training examples with such large objects we expect performance to increase with a larger database or dedicated set of training examples of pacemakers. Other types of foreign objects that were not or incompletely segmented were semi-opaque objects such as pens with plastic parts, hair, and certain types of necklaces. A good example is the case for glasses in Fig. 2.8, 2nd case. The glasses were removed well, but the case itself remains visible. Also here we hypothesize that the reduced performance is partially caused by a limited number of similar training examples in the database. Semi-transparent objects are also more difficult to precisely delineate manually because of their lower contrast with the background. We expect though that these objects will have a relatively small disturbing effect on automated analysis as they occur less frequently and their appearance is less conspicuous.

At the probability cut-off used for segmentation a false positive is detected in approximately one out of every 4 images (0.25 FP/image). Most of these false positives occur at the lung boundary, especially the interface of the lung with the chest wall (examples in Fig. 2.8, 4th case and Fig. 2.9, 1st case). We expect that the majority of this category of false positives would be easy to remove by adding extra position features to the classification, for example based on the distance to the lung border. False positives at the medial sides (near the heart and mediastinum) of the lung fields, or in the lung fields rarely occurred. Objects smaller than 120 pixels, corresponding to an effective diameter of 1.35-2.35 mm, were excluded in the FROC analysis and the texture analysis experiment. These objects are small compared to most foreign objects and we determined that also removing these objects from the image reduced texture scores only by 0.5%.

Detection and segmentation of foreign objects was limited to areas inside automatically detected unobscured lung fields. Depending on the location and appearance, objects close to the boundary of the lung fields can cause under- or oversegmentation of the lung fields. To also detect objects in oversegmented lung fields the observer providing manual segmentations was instructed to also segment objects close to the boundary of the lung fields. Examples of this can be seen in the 3rd case in Fig. 2.7 where a button is present close to, but outside of, the right upper lung and also in the 4th case of the same figure where the two zipper sliders at the top are outlined in the reference but outside the lung fields. As detection performance was only determined inside the lung fields this button was not counted as a false negative. Our main focus was to improve automatic analysis of the lung fields, therefore we limited the evaluation to this area. In some

cases object close to, but outside, the automatic lung segmentation might disturb the texture analysis inside the lung fields. Although we expect the effect on the texture scores to be small, the algorithm could be easily extended to include all foreign objects in the radiograph.

In general, inpainting results were visually convincing. Poor results sometimes occurred when the object was not fully segmented. In these cases the missed part will not be removed, but also the inpainting of the properly segmented parts can be distorted as there is a risk of selecting patches for inpainting from the missed part. In some cases sharp edges and ridges, such as those caused by ribs, are not fully restored. The method used in this paper does not explicitly try to continue structural elements in the image before the addition of texture. It might be that other inpainting methods, such as those of Bertalmio et al. [126] would handle continuation of structures in a better way.

After restoration the texture scores of images containing foreign objects are similar to those of normal images. This indicates that the statistical properties of the affected areas, measured by the features used in the textural abnormality system, are similar to unaffected areas after removal. Using downstream image analysis algorithm performance, such as segmentation or CAD, is an indirect way to evaluate a detection and removal algorithm and visually the removal does not have to be perfect to provide results which improve subsequent analysis. This observation is also made by Lee et al. [122] who state that image artifacts after hair removal do not influence the segmentation of dermoscopic lesions. Evaluating a hair removal algorithm, Abbas et al. [131] used a similar approach as in this work, using segmentation performance and the effect on texture measures to determine the quality of the processed images.

Removal of detected foreign objects is based on a modified version of the texture synthesis algorithm described by Efros and Leung [123]. The modifications were aimed at increasing the speed of the algorithm, as the original algorithm is known to be extremely slow. Computation time was reduced from a few hours using the original algorithm to less than a minute with the modifications (C++ implementation on a single 3 Ghz core). Compared to previously published techniques for texture synthesis [123,141], we have provided a method adapted to the intrinsic characteristics of medical images, particularly radiographs, with low computational time, which is a paramount feature for the analysis of large amount of images generated in a screening setting. The speed increase can be beneficial when the algorithm is integrated into a CAD system that is used to provide

feedback about image quality directly after acquisition of the radiograph. The full algorithm including feature calculation (15 seconds), pixel classification (4 minutes), segmentation (0 seconds) and inpainting (DullRazor 1 seconds, texture synthesis 30 seconds) takes approximately 10 minutes for an average case containing foreign objects. The majority of the computation time is spent on pixel classification using kNN. At a small loss of sensitivity, total computation times could be reduced to approximately 1 minute when SVM or LDA is used instead.

The different steps of the algorithm, such as pixel classification, segmentation or texture synthesis, require a number of parameter values to be set. Some of them, such as the scales to compute features on or the minimum size of objects to retain, were based on the general characteristics of the foreign objects. Parameter settings of classifiers are difficult or impossible to determine a priori and were selected by optimization on the training set. Others, such as the segmentation threshold, depend on the performance characteristics of previous computations and were based on a trade-off between over- and undersegmentation. The choice of object removal algorithm and the free parameters of texture synthesis were not extensively optimized in our application. The selected settings resulted in successfully removal of foreign objects from images, with texture scores indistinguishable from unaffected images. Further optimization of some parameters might be possible to improve detection and removal of particularly different foreign objects.

An alternative approach to the presence of artifacts in medical images is to ignore affected areas. Such an approach potentially misses pathological areas. Many CAD systems work with feature extractors that have non-local support, which requires the size of the excluded area to be larger than the size of the object itself and would further increase the chance of false negatives. Therefore, we believe that a detection and removal algorithm is preferable.

In radiological screening settings numbers of abnormal images are often low, in the order of a few percent. In the tuberculosis screening program in London, which was the source of the data used in this paper, the number of abnormal chest radiographs which needed referral for abnormalities compatible with active tuberculosis was on the order of 1%. If an automatic system were to be used to select cases not needing referral a considerable improvement in efficiency, by a reduction in workload and costs, could be achieved. The percentage of images containing foreign objects in this database was 20%. If a significant proportion of this 20% were selected as cases needing referral the efficiency improvement

would be much smaller.

The presented algorithm for detection and removal is general and can be applied to other objects in (chest) radiographs or to other modalities. In chest radiographs medical foreign objects, such as catheters and tubes, are one of the most common abnormal findings, accounting for 64% of the total in a study by MacMahon et al.[95]. In bedside radiographs, where these objects often occur, removing them can potentially improve the reading of these difficult low quality images by humans or automated systems.

2.7 Conclusion

An automated method to detect, segment and remove foreign objects on chest radiographs has been presented. The detection step is based on supervised pixel classification and evaluated using FROC analysis. The removal of the objects from the image is performed using texture synthesis inpainting. The effect of image restoration on false positive responses in a CAD system was determined in an experiment with a textural abnormality detection task.

We have found that high density foreign objects can be detected with high sensitivity with only a small number of false positives. The removal of the detected foreign objects from the image results in a reduction of false positive responses of a texture analysis system in normal images. This enables application of automated disease detection to improve efficiency of screening programs even with images of low quality due to the presence of foreign objects.

Acknowledgments

This study was supported by the European and Developing Countries Clinical Trials Partnership (EDCTP), the Evaluation of multiple novel and emerging technologies for TB diagnosis, in smear-negative and HIV-infected persons, in high burden countries (TB-NEAT) project.

We would like to acknowledge the work of Jane Knight and Diana Taubman, the two reporting radiographers on the mobile X-ray unit in London who collected all of the CXRs.

Clavicle segmentation

3

Laurens Hogeweg, Clara. I. Sánchez, Pim A. de Jong, Pragnya Maduskar, and Bram van Ginneken

Original title: Clavicle segmentation in chest radiographs

Published in: Medical Image Analysis 2012, 16(8):1490 – 1502

Abstract

Automated delineation of anatomical structures in chest radiographs is difficult due to superimposition of multiple structures. In this work an automated technique to segment the clavicles in posterior-anterior chest radiographs is presented in which three methods are combined.

Pixel classification is applied in two stages and separately for the interior, the border and the head of the clavicle. This is used as input for active shape model segmentation. Finally dynamic programming is employed with an optimized cost function that combines appearance information of the interior of the clavicle, the border, the head and shape information derived from the active shape model.

The method is compared with a number of previously described methods and with independent human observers on a large database, containing both normal and abnormal images. The mean contour distance with the reference contour of the proposed method on 249 test images is 1.1 ± 1.6 mm and the intersection over union is 0.86 ± 0.10.

3.1 Introduction

The automatic delineation of normal anatomical structures is a prerequisite for computerized analysis of medical images. The analysis of 2D radiographic images, such as chest radiographs, is a challenging task because superimposed normal and abnormal structures can make it difficult to discern the boundaries of particular objects. Detection, recognition and segmentation of these structures requires incorporating prior knowledge about their location and appearance. The large variation in both these properties inherent to medical imaging can be handled through the use of supervised systems that learn from examples.

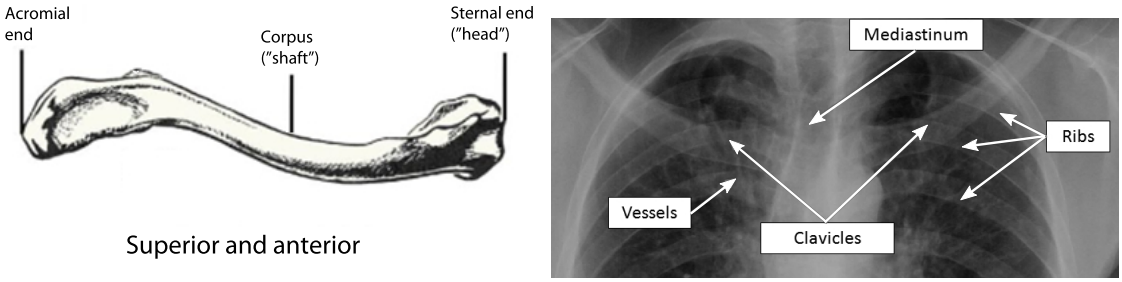
In this work we focus on the segmentation of the clavicles in chest radiographs. The clavicle is a cortical bone connecting the shoulder blade at the acromion to the breast bone at the sternoclavicular joint. Fig. 3.1(a) shows the anatomy of the clavicle in a schematic drawing. Fig. 3.1(b) shows the clavicle in a chest radiograph. The 2D projection in a radiograph causes several other structures to overlap with the clavicle. Notably these are the ribs, the mediastinum and the large vessels of the pulmonary vessel tree. In this paper the focus of the segmentation algorithm is the part of the clavicle contained inside the projection of the lung fields and the mediastinum. The lateral parts at the acromial end outside the lung fields are not considered.

Obtaining an accurate segmentation of the clavicles is useful for a number of applications. The segmentation can be used to digitally subtract the clavicle from the radiograph. Recently it has been shown that commercially available software that suppresses bony structures in the chest radiograph can improve the radiologist's performance[85,86]. Clavicle suppression in particular might aid radiologists to detect pathology in the lung apices, that are known to be difficult areas due to superimposing structures[151]. Performance improvement can also be expected if the automatic segmentation is used as input for computer-aided detection and diagnosis (CAD) systems. Certain lung diseases, such as tuberculosis, manifest themselves especially in the lung apex[111]. A good characterization of the structures in that area is needed to improve pathology detection and reduce false positives. Accurate localization of the medial parts of the clavicles can also serve to automatically determine possible rotation of the ribcage, an important quality aspect of chest radiographs. When chest radiographs are rotated, false abnormalities might appear in either or both of the lung fields due to apparent changes in parenchymal density.

Only a few papers have addressed the automatic segmentation of the clavicles. Yu et al.[152] used dynamic programming and a nonlinear shape model to segment the clavicles but no quantitative error analysis was performed. van Ginneken et al.[76] segmented the lung fields, heart and clavicles in chest radiographs from the JSRT database[37] and compared several segmentation techniques. While results comparable to the interobserver variability were obtained for the heart and lung fields, the segmentation of the clavicles proved a more difficult task. Seghers et al.[77] used a minimal shape and intensity cost path segmentation technique on the same database. The authors obtained a smaller error on the whole database than reported in van Ginneken et al.[76], but individual results for the clavicles were not provided. Simkó et al.[80] considered the task of detecting only the lateral part (diaphysis) of the clavicle. They attempted to find line shaped structures using the Radon transform on edge-enhanced images. The resulting segmentation was subsequently used to suppress the clavicles. Only a qualitative evaluation of the clavicle segmentation was provided.

The combination of different types of independently calculated information might help in the segmentation of anatomical structures, especially in difficult cases. We propose a method that combines different type of characteristics from the clavicles, such as border and head definition, local intensity and shape. These characteristics are extracted from the images by means of pixel classification (PC) and active shape models (ASM)[153] and combined using dynamic programming. The advantage of ASM is that it can only produce plausible shapes, unlike PC which treats segmentation as a local classification problem and can produce segmentations of any shape. In the case of confusing border information, such as overlying ribs or the presence of abnormalities, PC can produce shapes which do not resemble clavicles. A disadvantage of ASM is that its mechanism to generate plausible shapes (through principle component modeling of the training shapes) also limits how accurate it can outline the borders of a previously unseen instance of the structure. Especially when the shape is complex or has a large variability, such as in the case of clavicles, this problem becomes pronounced. This motivated us to use ASM to provide a plausible but rough outline of the clavicle and use multiple dedicated pixel classifiers to refine the outline. Subsequently an optimal cost path is calculated that ensures a globally optimal solution and provides a convenient framework to combine shape information and multiple pixel classifiers.

This paper is organized as follows. In Section 3.2 the data that was used



(a) Schematic drawing (b) Clavicles in the chest radiograph

Figure 3.1: Anatomy of the clavicle. (a) Schematic drawing of a clavicle with the main parts labeled. (b) Appearance of clavicles in chest x-ray with surrounding structures indicated. The corpus and sternal end will be colloquially referred to as respectively the shaft and the head, respectively.

for both training and evaluation of the system is described. The proposed new method for clavicle segmentation is detailed in Section 3.3, and various other methods to which the new method is compared are also described there. Section 3.4 and 3.5 presents a number of experiments and results, and these results are discussed in Section 3.6. Section 3.7 concludes.

3.2 Data

A set of 548 consecutively obtained posterior-anterior chest radiograph were selected from a database containing images acquired at two sites in sub Saharan Africa with a high tuberculosis incidence. All subjects were 15 years or older. Images from digital chest radiography units were used (Delft Imaging Systems, The Netherlands) of varying resolutions, with a typical resolution of 1800×2000 pixels, the pixel size was 250 μm isotropic.

From the total set of images, 225 were considered to be normal by an expert radiologist, while 323 of the images contained abnormalities. Of the abnormal images, 101 contained abnormalities in the area obscured by the clavicle. The data was divided into a training and a test set. The training set consisted of 299 images, the test set of 249 images. The development and optimization of the method was completely performed on the training set alone, the test set was only used to calculate the final results.

3.2.1 Manual segmentation of the clavicles

Manual tracings of the clavicle were used as the reference standard in this study. The clavicle is a high density object projected over a low density background (the

lung parenchyma). The border of the clavicle typically appears as a ridge with a higher density than the inside. Even for human experts it can be difficult to delineate this border precisely. The complex cross sectional shape of the clavicle causes multiple shadows on a chest radiograph. Often the border of the clavicle is partly aligned with the ribs. Especially the medial part of the clavicle is difficult to trace, because of multiple overlapping shadows from the vena cava, ribs, and mediastinal structures. The sternoclavicular joint is not always projected within the lung fields and can be hidden or very difficult to see.

Fixed points were used to determine the shape model for ASM and to evaluate different parts of the clavicle. The fixed points define three parts of the border of the clavicle: (1) the lower border, from fixed point 0 to 1, (2) the head from fixed point 1 to 2 and (3) the upper rib border, from fixed point 2 to 3. The following instructions to outline the clavicles were provided to human observers who provided manual tracings:

1. Start at the lateral inferior border of the clavicle at the projected crossing of the superior border of the scapula and the clavicle (fixed point 0)

2. Follow the inferior border until the start of the head (fixed point 1). The start of the head is defined as the location where the curvature of the border suddenly changes.

3. Follow the border until the end of the head (fixed point 2). The end of the head is defined as the location where the curvature of the border suddenly changes on the superior border.

4. Finish the segmentation by following the superior border until the projected crossing of the superior border of the scapula and the clavicle (fixed point 3).

The clavicles were outlined by three trained readers who were instructed by an expert radiologist. One of the readers was used to set the reference standard. The outlines of the other two readers are compared with the reference standard in the same manner as the automatic methods. These readers are referred to as the 2nd and 3rd observer. To ensure optimal outlines a regular review of the outlines was performed by the expert radiologist.

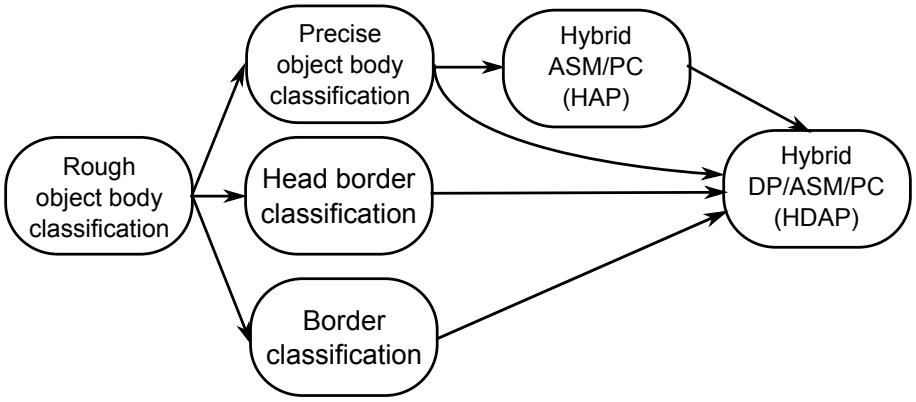


Figure 3.2: Flowchart of HDAP. A number of independent dedicated pixel classifiers are combined using active shape modeling (ASM) and dynamic programming.

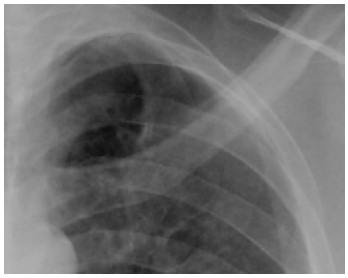
3.2.2 Public dataset and challenge

The database used in the paper is made publicly available so that it can be used by other research groups to evaluate their segmentation methods. The release of the images and annotations is part of our ongoing effort to improve the quality and progress of medical image analysis research by enabling fair comparisons of algorithms through public datasets and challenges. The dataset can be found at the website of the challenge Chest Radiograph Anatomical Structure Segmentation (CRASS) (`http://crass12.grand-challenge.org`). The chest radiographs and manual outlines of the clavicles of the training set are provided. For the test set only the radiographs are provided and the outlines are kept secret. New segmentation results on the test set can be uploaded to the website and will be evaluated automatically. Submitted results are automatically evaluated using similar measures as reported in this paper. An automated report and ranking among other methods will be publicly available for each submitted result.

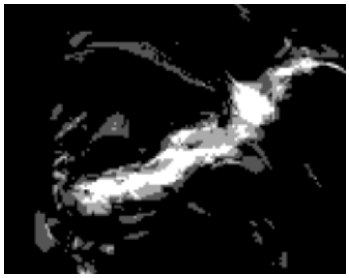
3.3 Methods

The proposed Hybrid Dynamic Programming/Active Shape Model/Pixel Classification algorithm (HDAP) combines a selection of existing methods in a structured way to improve on the results of the individual algorithms. A set of dedicated pixel classifier systems form the basis of HDAP, active shape modeling (ASM) is used to generate plausible shapes, and dynamic programming is used

to find the exact boundary. Each of the individual algorithms use the output of the previous step(s) in the algorithm as their input(s). A diagrammatic overview of the method is given in Fig. 3.2. The method is illustrated for one case in Fig. 3.3.

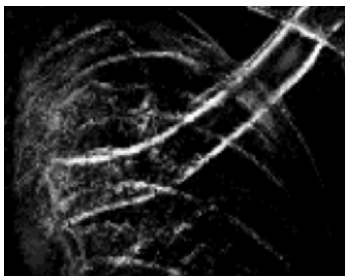


(a) Original X-ray



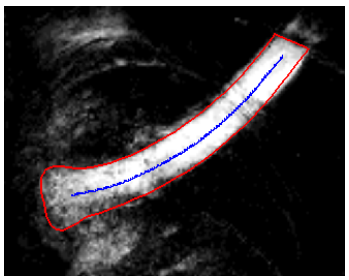
(b) Rough body classification (localization)

(c) Object body classification

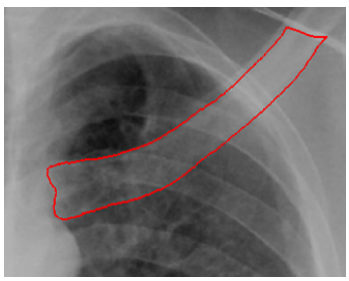


(d) Border classification

(e) Head border classification



(f) ASM fitted to object body



(g) Final segmentation produced by optimal path searching

Figure 3.3: Outputs of the algorithm shown for one case. Corresponding cost space images are shown in Fig. 3.4.

3.3.1 Pixel classification (PC)

Pixel classification forms the basis of the other methods. In this methodology the segmentation problem is recast into a pattern classification task[51,54]. A number of continuous characteristics (features) are calculated for a number of samples (positions, pixels) in an image. A classifier is trained using labeled samples from a

database of training images. Examples of labels are inside/outside clavicle and on/off the clavicle border. The classifier provides the mapping from features to class labels. In this work we use classifiers that provide a posterior probability that indicates the likelihood for a sample to receive a label. Test images can now be segmented by computing the features at each position and applying the classifier.

Features and classification

Three types of features were calculated for each sample: texture features based on Gaussian derivatives, features derived from the Hessian matrix and position features. First each image is resized to a width of 1024 pixels. To capture local image structure[133] the output of Gaussian derivative filtered images of order 0 through 2 (L, L_x, L_y, L_{xx}, L_{xy}, L_{yy}), at scales 1, 2, 4, 8, 16, 32 and 64 pixels were calculated. Hessian matrix derived features derived were used to detect the presence of line like structures[135]. Considering the two eigenvalues of the Hessian matrix λ_1, λ_2, $\lambda_1 > \lambda_2$ two measures were derived: (1) $\sqrt{(\lambda_1 - \lambda_2)^2}$ to extract the lineness of the local image structure and (2) the largest eigenvalue λ_1 to indicate the strength of the response. The typical location of the clavicles in the image was captured through a number of spatial features: the (x, y) coordinates in the image resized to a width of 1024 pixels, the normalized (x, y) coordinates inside the bounding box of the unobscured lung fields, the distance of the pixel to the boundary of the lungs, and the distance of the pixel to the center of gravity of both lungs. An automatic lung segmentation algorithm was used to find the unobscured lung fields[76].

In total 59 features were computed. To reduce computation time, features were sampled every 2^{nd} pixel on a regular grid in both the training and test images so that the resulting segmented images have a width of 512 pixels.

To construct the training set positive (clavicle) samples per image were randomly sampled (sample rates for each classification task are given below). For the negative samples the distance to the clavicle was used to control the sampling rate (see section 3.3.1). One sixth of the positive and negative samples was used for the final training set, while the remaining $5/6^{\text{th}}$ was used to evaluate the effect of feature selection and classifier selection.

Object body classification

A typical approach for segmentation using pixel classification is to detect all pixels inside the object[76]. Positive examples are selected from the inside of the object

Distance to border (pixels)	Class	Sampling rate
0-2	Positive	50% of available samples
2-3	-	not sampled
2-10	Negative	10% of available samples
10+	Negative	same number as nearby negative samples

Table 3.1: Sampling strategy for border classification

of interest and negative examples from outside. The clavicles are only a small part of the chest radiograph and to prevent the classifier focusing too much on the background structures the number of negative examples were sampled depending on the distance to the clavicle. 80% of the negative examples were sampled randomly within a distance of 10 mm to the clavicle, the remaining 20% was sampled from the rest of the image. A similar sampling strategy was successfully used to segment fissures in thoracic computed tomography scans[154]. The sampling rate for nearby negative and positive examples was 1.5%.

The computational burden of the classification was reduced by performing it in two steps. An initial rough classification of the object body was performed using a kNN classifier with $k = 5$ and the first 5 features selected by SFFS. This quick classification gives a robust initial detection and localization of the clavicles. A rectangular area around this approximate detection was determined to restrict the search space for the second classification stage. Posterior probabilities p were thresholded with $p = 0.5$, the bounding box was determined and dilated by 20 mm. Construction of the training set and classification of samples in test images was performed only in this search area for the subsequent precise pixel classifiers.

The output of pixel classifiers typically have a grainy noisy appearance. To reduce this effect and to obtain a single connected segmentation for each object (right and left clavicle) the output was post-processed. The output of the PC was blurred with $\sigma = 0.7$ mm and then thresholded with $p = 0.5$ to obtain a binary segmentation[155]. The two largest connected components were determined and holes were filled using a morphological closing with a circular kernel (radius = 10 pixels).

Object border classification

The border of the clavicles on chest radiographs has a distinct appearance from the body, appearing as a sharp dense ridge. In this work the border was separately classified from the object body to account for this difference in appearance.

The rationale is that samples, features and classifiers can be independently optimized for body and border separately. Positive samples, representing the border, were selected within a distance of $n = 2$ pixels from the outline of the clavicles. The negative samples were selected with a minimum distance of $1.5n$ pixels to the outline to prevent sampling false negatives as a consequence of the inaccuracy of manual outlining the border. To focus on the task of distinguishing the object border from its direct surroundings, 50% of the background pixels were selected within 10 pixels from the outline, the other 50% was randomly sampled within the previously determined approximate clavicle bounding box. The sampling rate for nearby negative and positive examples was respectively 10% and 50%. The sampling rules are summarized in Table 3.1.

Head border classification

The head of the clavicle is a very difficult area to segment automatically and has an appearance that is distinct from the shaft of the clavicle. A separate classifier was constructed for the head border. The sample strategy was the same as for the object border classification, but positive examples were taken only from the outline of the head (between fixed point 1 and 2). The sampling rate for nearby negative and positive examples was 30%.

3.3.2 Hybrid ASM/PC (HAP)

Active shape modeling (ASM)[153] is a popular method to segment structures in medical images. We use the implementation described in Cootes and Taylor[156] which uses a global shape model, a multi-resolution appearance model and a multi-resolution search algorithm. The steps of the algorithm are briefly repeated here.

The shape of one or more objects is described by n points combined in a vector $x = (x_1, y_1, ..., x_n, y_n)^T$. A shape model is trained from a set of training shapes by determining the mean shape and the principal modes of variations using Principal Component Analysis (PCA). Gray value profiles are sampled perpendicular to the object border at each point of the shape. The first derivative of the profile is calculated and is normalized. The best position of a point during search is determined by minimizing the Mahalanobis distance of the derivative profile to the appearance model created from profiles of corresponding points in the training set. The fitting of the shape is performed in an iterative scheme where points are alternately moved to their optimal position according to the appearance model

and then projected on the shape model. This projection is performed using least square fitting with a bound on the maximal variation. A multi-resolution appearance model is used to prevent the search algorithm finding local optima.

ASM is run on the output of the object body classifier instead of the original gray values, combining pixel classification and ASM into a hybrid ASM/PC (HAP) algorithm. The probabilistic output of the pixel classifier is used directly as input for the ASM method, without the postprocessing described in Section 3.3.1.

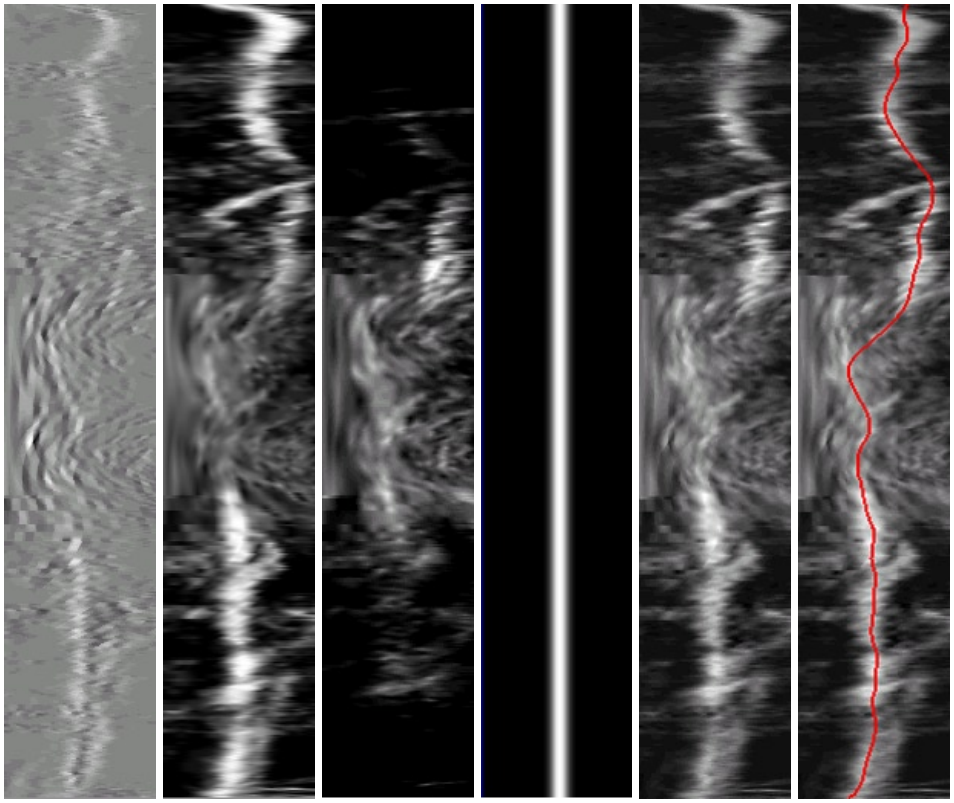
The ASM algorithm requires a set of training shapes with corresponding landmarks. The correspondence is provided by employing the four fixed points indicated during manual annotation (see Section 3.2.1). Between these fixed points a constant number of landmarks were interpolated over the initially annotated contour. On the lower clavicle border 20 points were interpolated, on the head 21 points, and on the upper clavicle border again 20 points, leading to a total of 65 landmarks per clavicle. The two clavicles are combined into one shape model containing two objects, consisting of 130 landmarks. By combining both clavicles in one model unlikely configurations, such as gross asymmetry, will be prevented during fitting as these configurations do not typically occur in (normal) chest radiographs.

3.3.3 Hybrid DP/ASM/PC (HDAP)

We note that the pixel classification method has high accuracy in areas where the border of the clavicles can be easily discerned. In areas where the output of the pixel classifier is uncertain another step is needed to integrate the local decisions. Contextual methods such as applying smoothing, mathematical morphology, iterated contextual classification[79] or Markov Random Fields[157] will generally improve the results by including information of the local surroundings but they do not provide a global integration of the available information.

Optimal path based methods[158,159], on the other hand, can provide this global context by finding the combination of local decisions that form the best solution given the evidence provided by the local (pixel classifiers) and global (shape) information sources. In an appropriate formulation the optimal path can be easily found using dynamic programming.

To detect the border of the clavicles with dynamic programming the image must be warped in an appropriate coordinate system. In the coordinate system we use here the border is an approximately straight line. The coordinate system



(a) Body cost (b) Object border cost (c) Head border cost (d) Shape cost (e) Total cost (f) Optimal path

Figure 3.4: Composition of cost function in optimal path space. High values indicate a high clavicle border likelihood. Figures (a-d) form the components of the cost function (see Equation 3.1). Figures (a-c) are derived from pixel classifiers by sampling profiles: (a) is based on Fig. 3.3(c), (b) is based on Fig. 3.3(d), (c) is based on Fig. 3.3(e). Figure (d) is the ASM fitted shape, based on 3.3(f). The total cost is shown in Figure (e), where (f) shows the optimal path.

is created by sampling profiles of pixel values from points on the border to the closest point at the primary medial axis (PMA, see Section 3.3.3) of the (estimated) object. The spatial sampling frequency is adjusted to the length of the profile so that in the transformed coordinate system the distance from PMA to the border is approximately the same amount of pixels. As dynamic programming must be able to improve the detected border, the profiles are extended on the outside of the object. A large margin of twice the PMA-contour distance was chosen as the length of the profiles. The result of this transformation is a rectangular optimal path space with one axis aligned to the clavicle border and the other axis aligned to the profiles.

Primary medial axis (PMA)

The skeleton of the output of the HAP method is found using the medial axis transform[160]. At the end of elongated objects the skeleton will typically have short branches emanating from the corners and running to the main centerline unless the ends are perfectly spherical. These (shorter) branches are removed until only one segment remains. The remaining single centerline of the elongated object is called the primary medial axis (PMA).

Cost function

Any type of image can be projected in the coordinate system of the optimal path space to form a cost function for the dynamic programming. The outputs of the different pixel classifiers described above and the output of HAP are combined in one cost function

$$C = C_{border} + \alpha C_{body} + \beta C_{shape}(\sigma) \tag{3.1}$$

where C is the final cost function, C_{border} the sum of the outputs of the object border and head border classification, C_{body} is the derivative of the output of the body classification in the profile direction (in optimal path space) and C_{shape} represents the influence of the shape found by HAP. The conversion of C_{body} by taking the derivative is needed to convert the original step edge formed by the border to a ridge in the cost function. The profile of C_{shape} is formed by Gaussian blurring (the scale controlled by the parameter σ) the clavicle outline that was obtained using HAP. The parameters α and β control the relative influence of each of the components. The construction of the cost function is illustrated in Fig. 3.4.

3.4 Experimental

The classifier settings, and the weights in the cost function combination were determined in a number of pilot experiments.

3.4.1 Feature and classifier selection

The optimal classifier for each of the pixel classifiers was determined using classifier and feature selection. The following classifiers were evaluated. A kNN-classifier with $k = 5, 15, 30, 50, 100,$ and 200 and a linear discriminant classifier (LDC) with and without Principal Component Analysis to reduce the number of features. Normalization of each feature to unit standard deviation and zero mean

was performed beforehand in all cases. For kNN classification the fast tree-based implementation by Arya et al.[136] was used. Previous work indicated that the approximate solution given by this implementation does not influence the classification results[76]. Feature selection was performed using Sequential Floating Forward Selection (SFFS)[143].

The effect of classifier and feature selection was evaluated using the area under the receiver operating characteristic (ROC) curve (A_z) on the pixel datasets. LDA classifiers performed in general much worse than kNN-classifiers. Minor differences in A_z were found between different values for k. Improvements in classifier performance (A_z) were not directly reflected in the segmentation performance measures (Section 3.5.1). A similar effect was found for the feature selection where small increases in classifier performance were observed with feature selection but no increase or even a decrease in segmentation performance. For this reason all subsequent results are shown for a kNN-classifier ($k = 15$) using the full set of 59 features and feature normalization.

3.4.2 Cost function weights

C_{border} and C_{body} in Equation 3.1 both indicate the border likelihood with the same scaling, the factor α controlling their combination was therefore set to 1. The influence of the shape model, determined by β and σ, was optimized by doing a grid search on these parameters in the training set. For β the values $(0.0, 0.05, 0.10, 0.20, 0.5, 1.0)$ and for σ the values $(1, 2, 4, 6, 8, 12)$ were tested. On the training set the combination $(\beta, \sigma) = (0.1, 4.0)$ yielded the best results and these values were used in subsequent experiments.

3.5 Results

3.5.1 Segmentation performance metric

To measure the performance of a segmentation algorithm, a 'goodness' index is required. For a two class segmentation problem, one can distinguish true positive (TP) area (correctly classified as object), false positive (FP) area (classified as object, but in fact background), false negative (FN) area (classified as background, but in fact object), and true negative (TN) area (correctly classified as background). From these values, measures such as accuracy, sensitivity, specificity, kappa and overlap can be computed. In this work we use the intersection divided by union as an overlap measure, given by

$$\Omega = \frac{TP}{TP + FP + FN}. \tag{3.2}$$

This is a well accepted measure, but one should be aware that objects that are small or thin or have a complex shape usually achieve a lower Ω than larger and more spherical objects[161]. For many purposes the part of the clavicle inside the lung fields is most relevant, and the overlap was calculated only inside the convex hull of the automatically segmented lung fields obtained using[76]. In addition, the mean absolute contour distance (MCD) is computed. For each point on contour X, the closest point on contour Y is computed; these values are averaged over all points; this is repeated with contours X and Y interchanged to make the measure symmetric[161]. The distances are given in millimeter. One pixel corresponds to approximately 0.85 mm on the images with a width of 512 pixels. For comparisons between methods, paired Student's t-tests to test the difference between means were used. Differences are considered significant if $p < 0.05$.

3.5.2 Segmentation results

The proposed HDAP method was compared with a number of other systems. Performance measures were calculated by considering the annotations of the first observer as the reference standard. The annotations of the 2$^{\text{nd}}$ and 3$^{\text{rd}}$ observer, the tuned version of the ASM from[76], as well as the object body classification with post-processing (PC-postproc) and the hybrid ASM/PC (HAP) method were evaluated. For ASM the same configuration as for HAP was used, except original gray values instead of posterior probabilities were used as input. In addition four different configurations of HDAP were evaluated to study the effects of including and excluding various components of the algorithm. 'HDAP' is the complete algorithm as described in Section 3.3.3. 'HDAP: no border' only uses shape and the object body classification to create the cost function in Eq. 3.1. 'HDAP: no head' does not include the dedicated head border classification in the cost space but does use the object border classification. The 'HDAP: no shape' variant does not include the term in Eq. 3.1 that controls the influence of the contour found by HAP.

Results of the different evaluated methods are shown in Tables 3.2 and 3.3 for respectively MCD and Ω. In Figures 3.5 and 3.6 the same results are shown graphically as boxplots[150]. In both figures and tables the results are ordered according to the median of the results (best performing first). Methods that show significant improvement compared to the previously listed method are indicated

MCD	mean	stdev	min	q1	median	q3	max
3rd observer	0.48	0.28	0.19	0.32	0.41	0.54	2.48
2nd observer *	0.49	0.25	0.19	0.34	0.43	0.57	1.94
HDAP: no shape *	1.09	1.57	0.28	0.57	0.81	1.15	22.74
HDAP *	1.10	1.57	0.27	0.56	0.82	1.16	22.65
HDAP: no head *	1.15	1.63	0.25	0.60	0.86	1.25	23.53
HDAP: no border *	1.49	1.58	0.39	0.92	1.20	1.59	21.44
HAP *	1.83	1.62	0.56	1.20	1.56	2.01	22.04
ASM	3.62	7.78	0.55	1.35	1.79	2.60	87.57
PC-postproc	2.74	4.46	0.94	1.50	1.86	2.47	44.28

Table 3.2: Segmentation results for the different methods. The mean contour distance (MCD; in mm) is given. Methods are ranked according to their median MCD. Methods which significantly ($p < 0.05$) improve over the previously listed method are indicated with an asterisk.

Ω	mean	stdev	min	q1	median	q3	max
3rd observer	0.93	0.04	0.70	0.92	0.94	0.95	0.97
2nd observer *	0.93	0.04	0.73	0.92	0.94	0.95	0.97
HDAP: no shape *	0.86	0.10	0.03	0.83	0.88	0.91	0.96
HDAP *	0.85	0.10	0.03	0.83	0.87	0.91	0.95
HDAP: no head *	0.85	0.10	0.02	0.82	0.87	0.91	0.96
HDAP: no border *	0.80	0.10	0.05	0.77	0.82	0.86	0.93
HAP *	0.77	0.10	0.05	0.74	0.78	0.83	0.92
PC-postproc *	0.73	0.11	0.18	0.70	0.75	0.80	0.89
ASM	0.69	0.19	0.00	0.65	0.75	0.80	0.90

Table 3.3: Segmentation results for the different methods. The overlap Ω is given. Methods are ranked according to their median Ω. Methods which significantly ($p < 0.05$) improve over the previously listed method are indicated with an asterisk.

with an asterisk.

All the HDAP methods improve significantly compared to the other automatic methods measured by both MCD and overlap. From the HDAP methods the no shape variant is the best performing method, with a slightly higher MCD and overlap than HDAP. HDAP: no border is the least performing of the variants, indicating that object border classification is an important part of the algorithm. When the head border classification is included in the cost function results improve significantly, especially the MCD.

Fig. 3.7 shows the outlines provided by the reference standard and five compared segmentations (2nd observer, HDAP: no shape, HAP, ASM and PC-postproc) for four selected cases. The four cases are chosen by ranking all results according

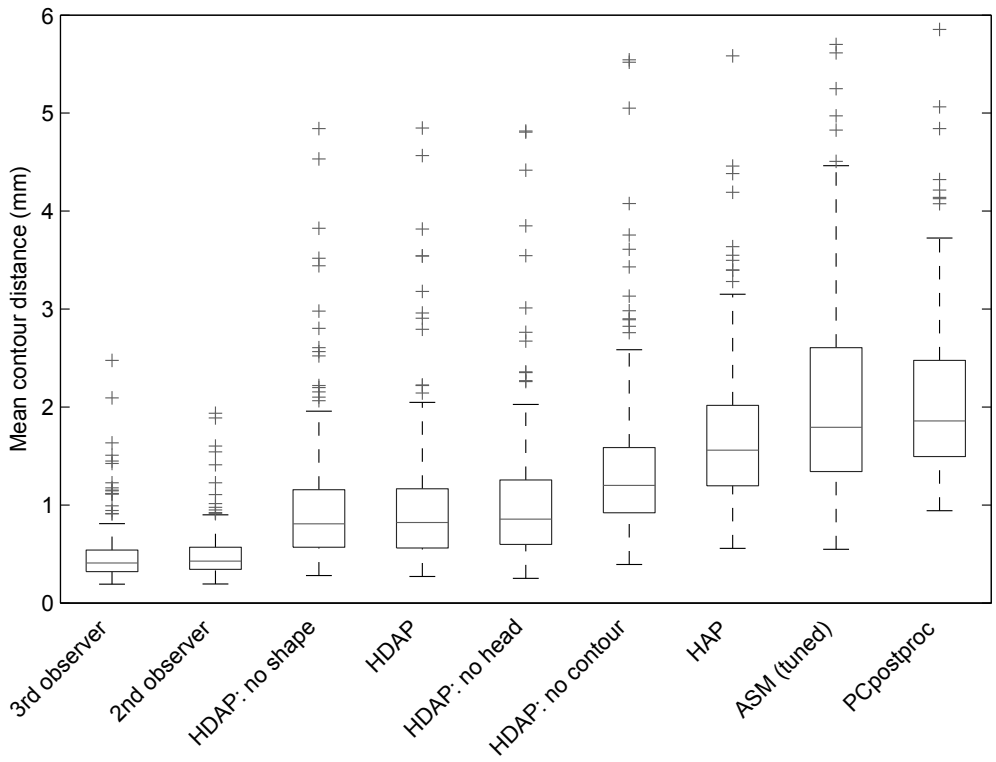


Figure 3.5: Boxplots of the mean contour distance on the test set of 249 cases for the different methods. The central line indicates the median, the box edges the 25th and 75th percentiles and the whiskers the extremes of the data excluding outliers. Points are considered outliers if they lie more than 2.7σ from the mean, corresponding to 99.3% percent of the data. The corresponding numbers are listed in Table 3.2.

to the average MCD of the four shown automated methods and then choosing for display the 0%, 33%, 66% and 100% percentile, respectively position #1, #85, #168 and #249. The performance increase as a result of the border refinement from HDAP to HAP can be clearly seen for all the cases. The last, worst segmented case, exemplifies one of the advantages of using HAP instead of ASM. In this radiograph the lung tops have been cut off due to poor collimation. ASM is not robust against these kinds of outliers resulting in 0 overlap with the reference standard, while the object body classification provides enough information for HAP to provide a reasonable segmentation.

3.5.3 Evaluation of different parts of the clavicle

The lateral parts of the clavicles are often clearly visible and relatively easy to segment, while the medial part of the clavicle is often obscured by numerous other structures such as the mediastinum or large vessels. The first row of Table

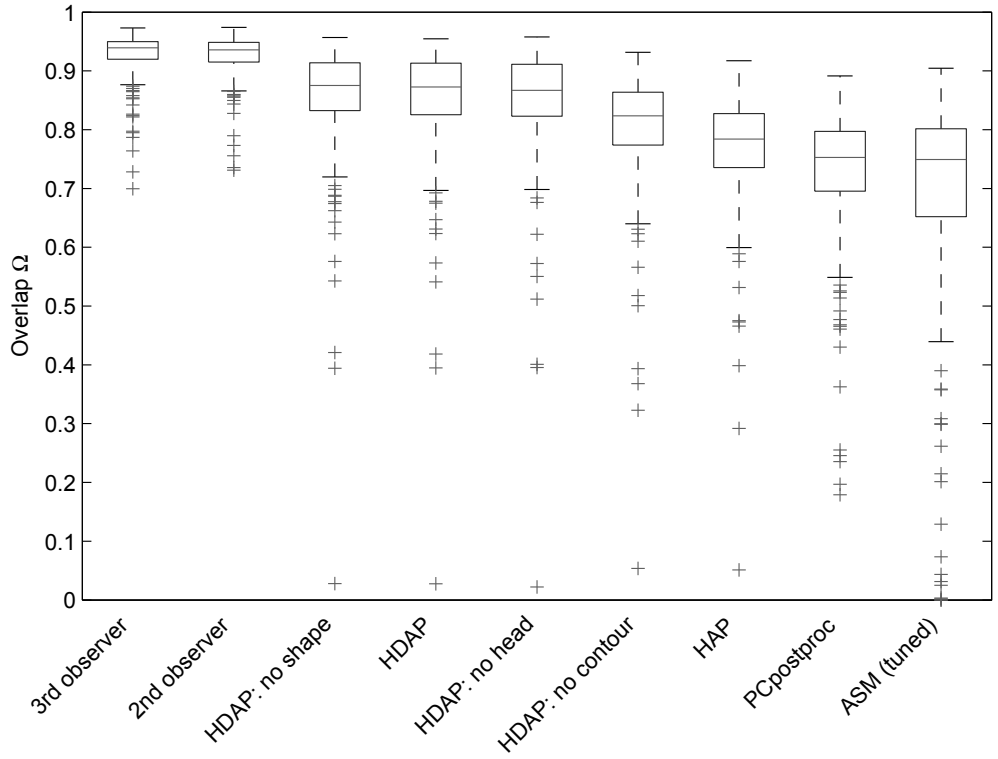


Figure 3.6: Boxplots of the overlap Ω on the test set of 249 cases for the different methods. The central line indicates the median, the box edges the 25th and 75th percentiles and the whiskers the extremes of the data excluding outliers. Points are considered outliers if they lie more than 2.7σ from the mean, corresponding to 99.3% percent of the data. The corresponding numbers are listed in Table 3.2.

3.4 shows for the best performing method, HDAP: no shape, how the algorithm performs for different parts of the clavicle. Shown is the mean MCD of all the test cases for 3 different parts of the clavicle (see Section 3.2.1).

The lower and upper border are most accurately segmented with an average error of approximately 0.5 mm. The head section of the clavicle is much harder to segment and has an error of about 2.5 pixels. To see if the adding of the head border classifier improved the results, the second row shows the results for the HDAP: no head method. The addition of the head border classification improves the MCD on the head by about 0.16 mm (0.2 pixels) on average. Results for HAP are shown in the last row and indicate that the addition of dynamic programming improves result by approximately 0.5 pixels for the lower and upper border and by 0.25 pixels for the head.

The bottom part of the table shows the same results for the abnormal subset of images in the dataset. The mean MCD for the lower and upper border is only

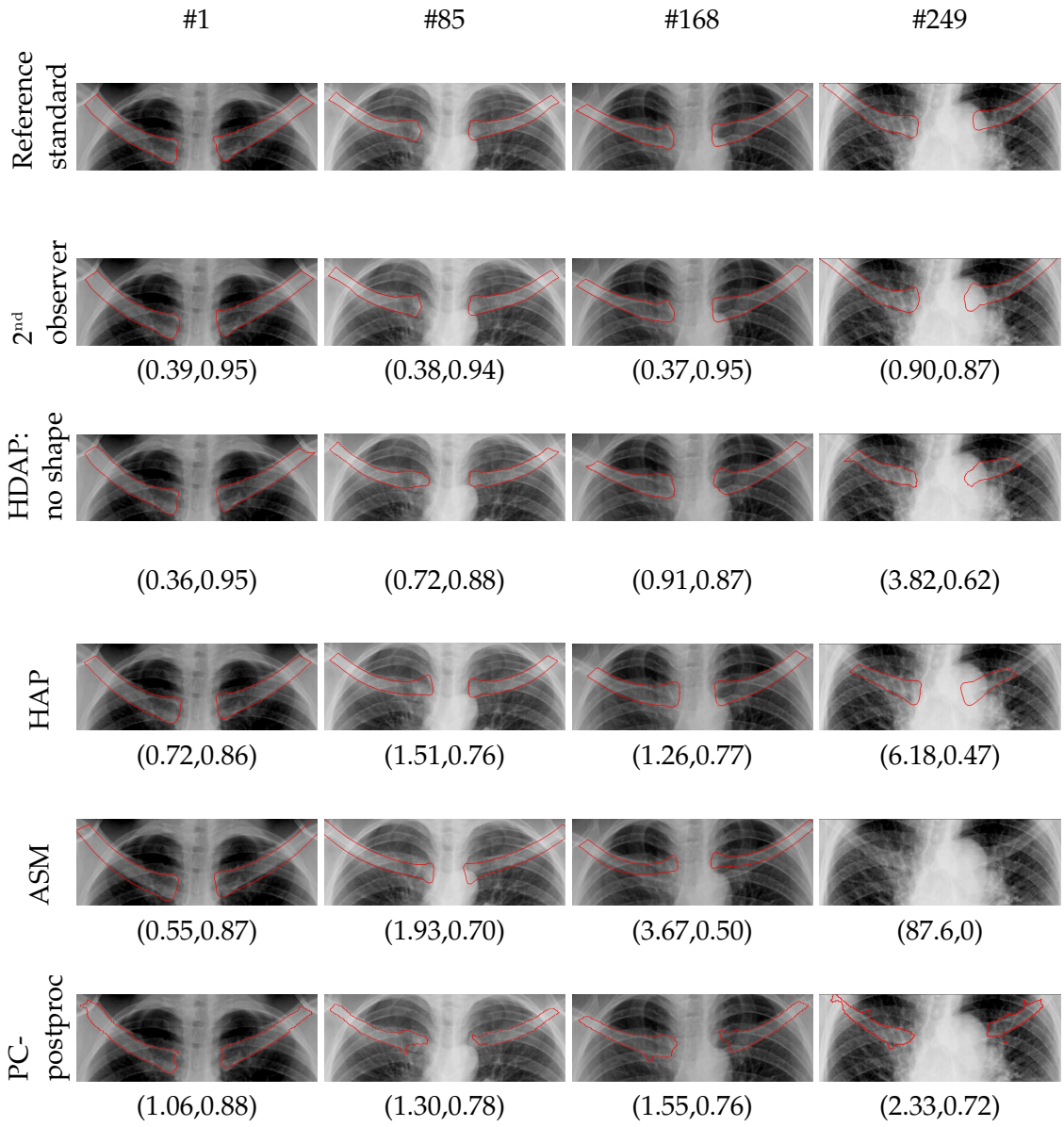


Figure 3.7: Examples for selected methods. Shown from left to right are respectively the #1, #85, #168 and #249 cases ranked to the average MCD of the displayed methods (from top to bottom). Indicated below each example are the MCD (in mm) and the overlap Ω.

slightly higher compared to the results for all images, while errors increase more in the head.

3.5.4 Effects of ASM and DP

Fig. 3.8 shows the effects of the different components of HDAP on the MCD. Each point in the plot is a radiograph, if a point lies on the identity line no improve-

	All images		
Method	Lower border	Head	Upper border
HDAP: no shape	0.49±1.27	2.84±3.33	0.61±1.64
HDAP: no head	0.49±1.24	3.00±3.33	0.59±1.61
HAP	1.18±1.15	3.39±3.06	1.08±1.72
	Abnormal images		
	Lower border	Head	Upper border
HDAP: no shape	0.60±1.79	3.56±4.36	0.63±1.50
HDAP: no head	0.60±1.76	3.72±4.31	0.61±1.41
HAP	1.27±1.43	4.03±3.97	1.10±1.40

Table 3.4: Mean contour distance (MCD; in mm) on the different sections of the clavicle. Values given are mean ± standard deviation. The first row shows the best performing method. The second row indicates the performance difference when the head border cost function is not used. For comparison the results of HAP are also shown.

ment was observed between the two compared methods, if the point is above the line the error has increased, if it is below the line the error has decreased. Fig. 3.8(a) shows the change of error when ASM is applied to the pixel classification output (HAP). In most cases the error decreases, indicating the ability of HAP to correct for some of the errors made by PC-postproc. A number of cases show a large reduction in the error. Fig. 3.8(b) shows the same type of scatter plot for the change in error dynamic programming is added (from HAP to HDAP). For the large majority of cases the error decreases again. Cases with a large error when using HAP also typically have a large error after applying HDAP, indicating that HDAP especially improves on cases where the true border has been found already approximately.

3.6 Discussion

A hybrid segmentation algorithm, HDAP, has been presented to segment the clavicles in chest radiographs and has been evaluated on a large database of normal and and abnormal radiographs. Two main conclusions can be drawn from the results: (1) the addition of dynamic programming to a combination of other algorithms significantly improves the segmentation performance compared to the previous state-of-the-art algorithm, and (2) the automated segmentation of clavicles in chest radiographs is a difficult problem and does not yet achieve the same performance as human observers. In this discussion the merits of the presented HDAP algorithm will be pointed out first, including a comparison with previously published work. Then possible reasons for the lower performance

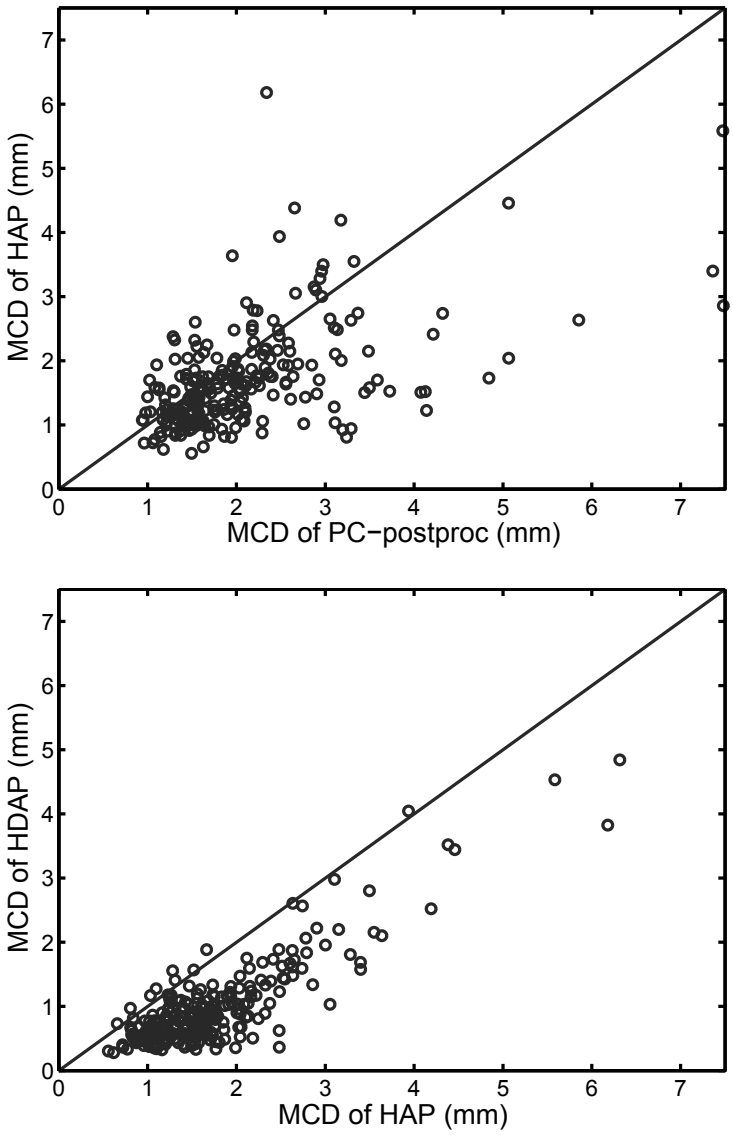


Figure 3.8: Effect of steps in HDAP on MCD. The left scatterplot shows the change in MCD from the output of the post-processed body pixel classification (PC-postproc) to HAP, each point being one case. For most cases an improvement of the MCD is observed. The right image shows a similar plot for the change from HAP to HDAP, in almost all the cases an improvement in MCD occurs.

compared to human readers are discussed. Finally a number of recommendations for future research are given.

3.6.1 Optimal path searching as contour refinement

Searching for an optimal path in a cost space in order to segment objects is an approach with some favorable properties. By modifying the cost space in an appropriate way it is easy to encode extra information into the algorithm[162]. The

use of dynamic programming to locate the optimal path ensures a global solution to the problem. HDAP provides a double integration of local and global information. First local cues provided by the object body classification are integrated with shape knowledge using HAP and subsequently additional local cues provided by different pixel classifiers are added. For the sake of simplicity and elegance it might be argued that one should focus on using and optimizing one individual algorithm instead of combining an array of methods. It is unlikely though that one single algorithm can robustly deal with all kinds of variation encountered in medical image segmentation.

The basic problem in all combinations of methods is to retain the good parts of each method and discard the bad parts. In the case of segmentation it can be expected that certain algorithms perform better at localizing certain parts of the object or perform better for certain cases. In general it is very difficult to determine for a single specific case whether a method has succeeded (at specific locations) or not. If a procedure would exist to detect errors in method for particular cases it is often also possible to correct these errors. A few of such attempts have been performed. In an application to segment breast mass on ultrasound Cui et al.[163] defined a goodness-of-fit criterion for points on a contour found using a snake model. If the points were not properly located, a second specialized fitting stage was used to improve the segmentation. van Rikxoort et al.[164] used error detection based on shape statistics derived from a training set of failed segmentation results to switch to a more advanced (and computationally more expensive) segmentation algorithm in case of a failure. In the general case, when such a procedure does not exist, some sort of averaging or majority voting can be used[76,165,166].

We argue that HDAP uses a special kind of averaging where evidence is accumulated from different types of pixel classifiers by adding them in the cost space. In general increasing the number of components in a system is expected to increase robustness as long as the errors of the individual components are not too large and are complementary to each other, a result well known from the field of classifier combination[167]. For this particular application of segmenting the clavicles especially detecting the border separately of the object proved to yield a large performance improvement.

The idea of refining an initial rough object detection has been explored before in[168]. A shape variant Hough transform was used to solve the problem of generating an initial detection and outline of the object of interest. A border appearance model was then created by sampling profiles perpendicular to the border and cal-

Method	van Ginneken et al.[76]	This work
ASM	2.04	3.52
Hybrid ASM/PC (HAP)	2.78	1.60
PC-postproc	2.90	1.92
HDAP: no shape	-	0.89
2nd observer	0.68	0.44

Table 3.5: Comparison of results in this paper with van Ginneken et al.[76]. For a fair comparison only images containing no abnormalities in the lung top (148/249) are used to calculate the mean MCD.

culating a number of features from them. The fit values of border points (based on Mahalanobis distance) in a test image is then used in an active snake model to refine the initial outline. While similar in overall approach of the problem there are some important differences with our work. The initial detection of HDAP is based on the appearance of the texture of the object and not only on the appearance of the border. This adds improved robustness when the border detection provides confusing results in individual images or when it is difficult to generate an accurate border appearance model. The generation of the cost space for border refinement also differs; HDAP uses cost functions derived from a set of dedicated pixel classifiers which can be easily expanded by choosing different training sets for different sections.

3.6.2 Comparison with previous approaches to clavicle segmentation

In previous work on the segmentation of clavicles in chest radiographs van Ginneken et al.[76] the set-up for the ASM was different as also the heart and the lungs were included in the model. Adding the lungs to the shape model might give a benefit in some cases, but due to the inherent limitation of flexibility of shape models it is not expected that performance will actually increase. The database used in this work is also different from the JSRT database used in van Ginneken et al.[76]. The JSRT database is a lung nodule database and in only a few cases the clavicular area is affected by the presence of a nodule. Instead, the database used in this work contained a considerable number of abnormal images.

To be able to compare the results, Table 3.5 shows results for the images containing no abnormalities in the lung top. The mean MCD for ASM is higher in this work than in van Ginneken et al.[76]. The use of (severely) abnormal training images can be an explanation for this. Also the "tuned" parameters for ASM

were actually tuned on the database from van Ginneken et al.[76]. These issues were not further investigated because ASM only serves as a reference method. HAP shows higher performance than ASM here, while on the JSRT database the reverse is observed. This change can be a consequence of the higher robustness of HAP against pathology and the better performance of the object body classification. Most importantly, HDAP (all variants including no shape, the best one), shows errors smaller than the other methods on either of the two databases.

HDAP shows some similarities to the method proposed by Yu et al.[152]. These authors used a method that alternates the application of nonlinear shape model fitting and dynamic programming contour refinement to the segmentation of both lungs and clavicles in chest radiographs. While they stress the importance of nonlinear shape models compared to standard PCA, their method fails to improve over standard ASM for the lung fields. A reason for this could be that the cost function in their search space is based only on intensity differences (to find the edges) instead of on multiple specialized border detectors as in HDAP. No numerical results are given in Yu et al.[152] for the accuracy of the segmentation of the clavicles.

Simkó et al.[80] developed a method to detect the diaphysis (shaft) of the clavicle using an initial detection with a Radon transform and a subsequent contour refinement using active contour fitting. Only qualitative results on the segmentation accuracy were reported. These authors also showed the ability of suppressing the previously automatically detected clavicle to reduce the number of false positives of a nodule detection system applied to the publicly available JSRT database[37].

Seghers et al.[77] used a combination of local (point) appearance models and local shape information based on the orientation vector between two consecutive points to generate a cost space. By using dynamic programming to find the optimal path in this space, the problem of generating plausible shapes with control points at plausible locations is solved at once. This method, called minimum intensity and shape cost path (MISCP) has the important advantage that no iterated methods, such as ASM, which suffer from local minima, are used. The immediate drawback of this is that no global shape information is encoded and that the method might produce unrealistic results when individual points can not be detected reliably. Using the PC output as input for ASM in HDAP largely solves the problem of local mimima during fitting. Also the border refinement in HDAP ensures a precise segmentation over the whole border of the object instead of only

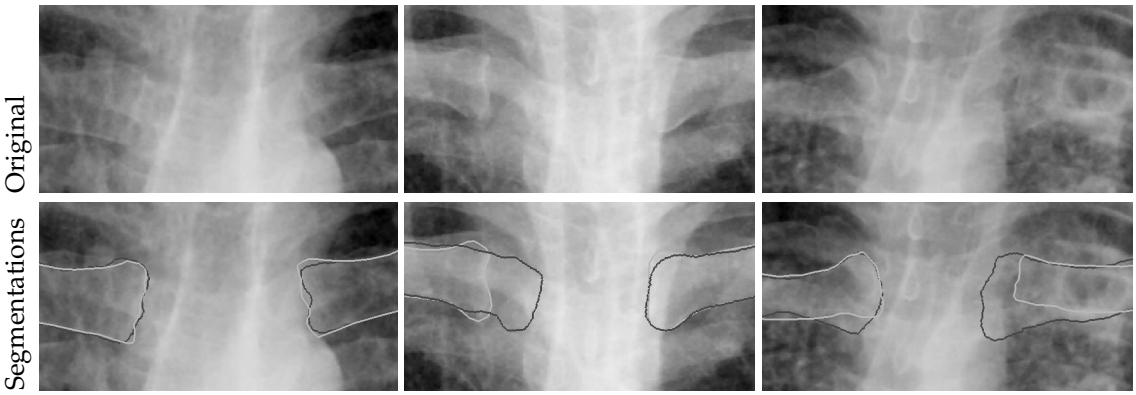
at a number of fixed points as in MISCP.

3.6.3 Improving HDAP

The mostly linear relation between the MCD of HAP and HDAP (Fig. 3.8) indicates that if the initial border is located too far away from the real border, HDAP cannot improve on HAP. The maximal correction that HDAP can achieve is determined by the length of the profiles in optimal path space. This length was set to double the original distance between border and centerline (PMA) of the object, but was not optimized. If the initial pixel classification contained very large errors, HAP cannot improve over PC-postproc.

The slightly higher MCD on the upper clavicle border (see Table 3.4) can be explained by the presence of a double ridge which is often seen on the superior side of the clavicle but not on the inferior. For human observers it is relatively easy to choose the correct line if the clavicle border is traced from medial to lateral. HDAP sometimes chooses the wrong line when there is a more optimal path for the dynamic programming. This problem could be partly solved by adding a term to the cost function which prevents sudden changes in the smoothness of the optimal path. In practice, such corrections are typically not easily made without introducing errors in the segmentation of other images.

The error at the head of the clavicle is considerably larger than the errors at the diaphysis of the clavicle. In some cases HDAP can locate the border of the clavicle at the head very precisely (Fig. 3.9(a)); in other cases the algorithm fails even when a clearly visible edge can be seen (Fig. 3.9(b)). Finally in a considerable



(a) Accurate segmentation (b) Failure of HDAP (c) Difficult case

Figure 3.9: Examples of head border segmentation. The light colored line is the outline provided by the reference standard, the dark colored line from the best performing method HDAP: no shape.

part of the cases the medial edge of the clavicle is hidden behind the mediastinal structures and both HDAP and the human observers fail to agree on where the border should be (Fig. 3.9(c)). How to improve the performance of automated segmentation methods in such difficult areas as the head of the clavicle is an open question. Human observers likely include much more context information in their reasoning. This context knowledge is provided by general knowledge, specific domain knowledge and their ability to reason about that knowledge and weigh alternative options. Encoding a comprehensive part of this knowledge in a computer system has so far proven very difficult.

3.6.4 Computation time

HDAP was implemented on a 3 Ghz Intel Core 2 Duo with single threaded C++ code. The total computation time for one case is about 18 minutes. Most of the time is spent in the calculation of the features for the images (about 13 minutes), followed by the pixel classifiers (about 5 minutes). The combination of the information through HAP and HDAP costs only a few seconds of the total time. Running time is high but our implementation was not optimized. A more in-depth feature selection analysis could reduce the number of features that need to be calculated, e.g. Gaussian derivatives calculated at large scales are expensive to compute but mostly contribute to the localization of the clavicle and not to the exact determination of the border. A large speedup can also be obtained by using a recursive implementation of Gaussian derivatives[169]. Alternatively, efficiently computed filter families such as Haar wavelets could be used[170]. Speedup for pixel classification can be obtained by parallelizing the classification or by adopting faster classifiers with similar performance. Also a number of specific strategies can be adopted to reduce the number of samples that need to be classified, such as multi-resolution schemes[76] or sparse pixel classification[171].

3.7 Conclusion

A new method (HDAP) to automatically segment the clavicles in chest radiographs has been presented. HDAP combines three segmentation frameworks: pixel classification applied in two stages and separately for the interior, the border and the head of the clavicle, followed by active shape model segmentation and, finally, dynamic programming using an optimized cost function. The method is compared with a number of previously described state-of-the-art methods and simplified versions of the proposed method. The large database that was used

for training and testing the method is made publicly available to facilitate future comparisons with other methods. Results were analyzed quantitatively with a number of standard measures and compared to two independent human observers.

The main conclusion is that a combination of several existing techniques (multiple pixel classifications, active shape modeling and dynamic programming) leads to a robust algorithm that significantly outperforms previously proposed methods. Dynamic programming is a convenient way to combine information from a number of algorithm components. Yet, the problem cannot be considered solved, as the errors of the automatic methods are larger than the inter-observer variability. Still, we believe results are good enough, especially in normal images, to use the segmentation for subsequent steps such as clavicle suppression, measuring rotation of the rib cage and computer-aided detection of abnormalities in the lung apices.

Acknowledgements

The authors gratefully acknowledge Dr. H. Ayles of the ZAMBART Project (University of Zambia, Lusaka, Zambia) and the Department of Clinical Research, London School of Hygiene and Tropical medicine (London, UK) for kindly providing the chest radiographs and T. Dubbelink, N. Snellen and K. Hengstler for outlining the clavicles.

Suppression of translucent elongated structures

4

Laurens Hogeweg, Clara. I. Sánchez, and Bram van Ginneken

Original title: Suppression of translucent elongated structures: applications in chest radiography

Published in: IEEE, Transactions on Medical Imaging, in press

Abstract

Projection images, such as those routinely acquired in radiological practice, are difficult to analyze because multiple 3D structures superimpose at a single point in the 2D image. Removal of particular superimposed structures may improve interpretation of these images, both by humans and by computers. This work therefore presents a general method to isolate and suppress structures in 2D projection images.

The focus is on elongated structures, which allows an intensity model of a structure of interest to be extracted using local information only. The model is created from profiles sampled perpendicular to the structure. Profiles containing other structures are detected and removed to reduce the influence on the model. Subspace filtering, using blind source separation techniques, is applied to separate the structure to be suppressed from other structures. By subtracting the modeled structure from the original image a structure suppressed image is created.

The method is evaluated in four experiments. In the first experiment, ribs are suppressed in 20 artificial radiographs simulated from 3D lung computed tomography (CT) images. The proposed method with blind source separation and outlier detection shows superior suppression of ribs in simulated radiographs, compared to a simplified approach without these techniques. Additionally, the ability of three observers to discriminate between patches containing ribs and containing no ribs, as measured by the Area under the Receiver Operating Characteristic curve (AUC), reduced from 0.99-1.00 on original images to 0.75-0.84 on suppressed images. In the second experiment clavicles are suppressed in 253 chest radiographs. The effect of suppression on clavicle visibility is evaluated using the clavicle contrast and border response, showing a reduction of 78% and 34%, respectively. In the third experiment nodules extracted from CT were simulated close to the clavicles in 100 chest radiographs. It was found that after suppression contrast of the nodules was higher than of the clavicles (1.35 and 0.55, respectively) than on original images (1.83 and 2.46, respectively). In the fourth experiment catheters were suppressed in chest radiographs. The ability of three observers to discriminate between patches originating from 36 images with and 21 images without catheters, as measured by the AUC, reduced from 0.98-0.99 on original images to 0.64-0.74 on suppressed images.

We conclude that the presented method can markedly reduce the visibility of elongated structures in chest radiographs and shows potential to enhance diagnosis.

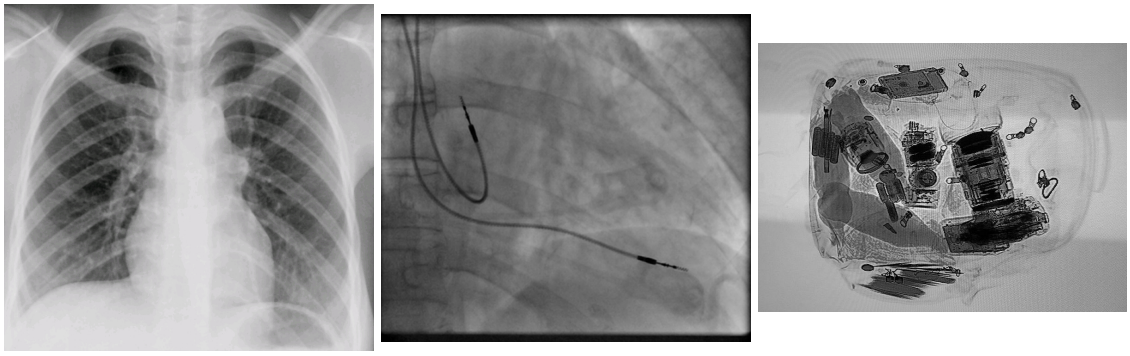


Figure 4.1: Examples of 2D projection images. Information about the 3rd dimension is lost in the acquisition process.

4.1 Introduction

Two-dimensional projection images are commonly made for a multitude of purposes and are daily acquired in large quantities in clinical radiology. Fig. 4.1 shows some examples of commonly acquired 2D projection images. An identifying property of these images is that multiple 3D structures are superimposed at a single point in the 2D image. This overlapping effect might partially obscure regions of interest in the image, reducing their visibility and making correct interpretation challenging for both humans and automated systems. Therefore, image processing methods aimed at identifying and removing the effect of superimposed structures in 2D projection images have the potential to reduce manual and computer-based interpretation errors.

Among medical projection images, the chest radiograph is the most commonly performed diagnostic exam in the world[31]. Chest radiography is widely applied to diagnose diseases such as tuberculosis, pneumonia, and lung cancer. These and other chest diseases are an important cause of mortality, leading to 10 million deaths annually[172]. A major difficulty in the manual and automatic reading of chest radiographs is the presence of superimposed normal structures such as ribs, clavicles, catheters, and vessels. These structures confuse interpretation and hide abnormalities, causing important decision-making errors[151,173].

Several studies have addressed the problem of analyzing chest radiographs where the lung fields are obscured by overlapping normal anatomy. Giger et al.[88] proposed a method to improve the detectability of nodules by suppressing the normal background using an image difference technique, in which a nodule-suppressed image is subtracted from a nodule-enhanced image. A similar filtering method that suppresses elongated objects (ribs) and enhances sphere-like

objects (nodules) before classifying nodule candidates was used by Keserci and Yoshida[89]. In both studies the individual effect of the filtering method on the performance was not reported. Chen et al.[92] classified automatically selected square regions in the lung using power spectrum based texture measures. Regions containing high gradient edges with an orientation corresponding to ribs were removed. A high classification performance of images affected by interstitial lung disease was reported. Loog and van Ginneken[91] presented a general filter framework based on regression, which has been applied to the suppression of bony structures on chest radiographs. The method gave promising results but was not further evaluated on a clinical problem. Suzuki et al[90] suppressed ribs using an artificial neural network and showed that this technique increased the visibility of nodules[174] and improved the quality of temporal subtraction images[175]. Simkó et al.[80] suppressed clavicles by creating a bone model from a gradient map smoothed along the clavicle border direction, after which a clavicle free image was created by subtraction of the model. They showed promising results of clavicle suppression on reducing false positives in a nodule detection task. Recently, it has been shown that suppression of bony structures in the chest radiograph can improve the radiologist's performance to detect nodules[85–87].

In this paper, we propose a method to remove unwanted structures on 2D projection images, particularly elongated structures. Elongated structures, like bones or tubes, are common in natural images, such as those acquired in medical imaging. The proposed method uses blind source separation techniques together with outlier identification to estimate an intensity model of the unwanted structures and subsequently remove them from the original image. Common artifact removal techniques[123,126,127] estimate the structure model by extracting information from image areas where the artifact is not present. In contrast, the proposed method does not require the presence of unaffected areas to remove the structure; our model is estimated using only the intrinsic properties of projected structures, such as elongated shape and translucency. The main goal of this study was to establish a general algorithm for suppressing elongated structures. We thoroughly evaluate the method in experiments where three different elongated structures commonly found in chest radiographs are suppressed, namely ribs, clavicles and catheters.

The paper is organized as follows. Section 4.2 describes isolation and suppression of elongated structures. Experiments and results are provided in Section 4.3. Discussion and conclusion are presented in Section 4.4 and 4.5.

4.2 Methods

In this section, we describe a general algorithm for the isolation and removal of an elongated structure of interest $\mathcal{S}$ in a 2D projection image from a background with other structures present. The goal is to estimate a projected image that is similar to the projection of the 3D scene in which the structure of interest was not present. To achieve this goal the image is decomposed into an image containing only the structure and one containing the background.

We assume that the 2D projected image $L(x, y)$ can be linearly decomposed into independent components as follows

$$L(x, y) = \sum_i L_i(x, y), \tag{4.1}$$

where $L_i(x, y)$ is the 2D image of one component, a structure of interest in the image. For a case with only one structure of interest $\mathcal{S}$, Eq. 4.1 can be written as

$$L(x, y) = L_{\mathcal{S}}(x, y) + L_{\mathcal{B}\setminus\mathcal{S}}(x, y) \tag{4.2}$$

where $L_{\mathcal{S}}(x, y)$ is the projection image of the structure $\mathcal{S}$ only and $L_{\mathcal{B}\setminus\mathcal{S}}(x, y)$ is the projection image of the 3D scene projected in $L(x, y)$ but without the presence of $\mathcal{S}$. To perform this decomposition two conditions are assumed to be fulfilled: (1) the structure $\mathcal{S}$ can be modeled in a 2D projection, and (2) the structure is translucent. The algorithm focuses on modeling of elongated structures. The specific appearance of these structures makes it possible to derive a model for $\mathcal{S}$ using the local intensity information only. The condition of translucency is expressed by the linearity of the composition, i.e. a fully translucent structure has no non-linear interactions with other structures in the imaging process.

To remove translucent elongated structures from natural images we propose two steps: (1) modeling and reconstruction of the elongated structure using subspace filtering, and (2) suppression of the identified structure from the original image. The next two sections describe these steps for one instance of an elongated structure $\mathcal{S}$. If multiple instances of elongated structures are present they can be modeled and removed in succession in order to obtain a final estimate $\hat{L}_{\mathcal{B}\setminus\mathcal{S}}(x, y)$ of $L_{\mathcal{B}\setminus\mathcal{S}}(x, y)$.

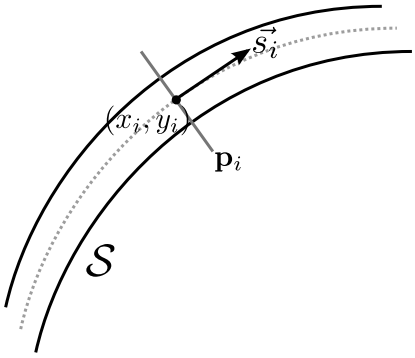


Figure 4.2: Definition of structure response. Given a structure $\mathcal{S}$ and an aligned curve γ, profiles $\mathbf{p}_i$ are sampled perpendicular to the direction $\vec{s_i}$ of γ at locations (x_i, y_i).

4.2.1 Modeling and reconstruction of the structure

The purpose of this step is to isolate/reconstruct the image response of $\mathcal{S}$ by filtering out the superimposed responses from other structures. Assuming that an observed intensity in the image is a mixture of unknown independent sources, we can use blind source separation (BSS) techniques to filter out the unwanted responses. Such an approach has been widely used, for example in wireless communication[176], speech processing[177] and EEG analysis[178].

Given a group of observations $\mathbf{X} = [\mathbf{x}_i \quad \mathbf{x}_{i+1} \quad ... \quad \mathbf{x}_n]$, composed of a mixture of underlying independent sources $\mathbf{Z}$, BSS techniques allow to recover an estimate of the sources $\hat{\mathbf{Z}}$ by identifying a demixing matrix $\mathbf{W}^{-1}$

$$\hat{\mathbf{Z}} = \mathbf{X}\mathbf{W}^{-1}. \tag{4.3}$$

Reducing the rank of $\hat{\mathbf{Z}}$, in such a manner that unwanted or uninteresting sources are removed, subspace filtering can be performed[178,179] and a filtered version $\mathbf{X}_F$ of the observations $\mathbf{X}$ is reconstructed

$$\mathbf{X}_F = \hat{\mathbf{Z}}_r\mathbf{W}. \tag{4.4}$$

The components in $\hat{\mathbf{Z}}_r$ form a local model of the structure, which is used to separate it from the background structures. In order to apply BSS to reconstruct the image response of the structure $\mathcal{S}$, we perform the following steps: (a) definition of the observed structure responses; (b) outlier detection; and (c) subspace filtering by means of BSS.

Observed structure responses

Given a curve segment γ, aligned with the structure $\mathcal{S}$, let $\mathbf{p}_i = \{p_{i1}, p_{i2}, ..., p_{iM}\}$ be an observed profile with length M, sampled through the point $s_i = (x_i, y_i)$ on γ and perpendicular to its direction $\vec{s_i}$ (Fig. 4.2). Intensity values of the profile are determined by linear interpolation from the original image. We represent the observed structure responses $\mathbf{P}$ as a group of N profiles evenly spaced along γ

$$\mathbf{P} = [\mathbf{p}_i \quad \mathbf{p}_i \quad ... \quad \mathbf{p}_N]. \tag{4.5}$$

$\mathbf{P}$ can be interpreted as an image patch, with dimensions $N \times M$, of the structure that has been straightened so that the cross sections of the structure align. To separate elongated structures in 2D projection images from their background, M and N must be set to values appropriate for the type of structure being suppressed. To ensure good subtraction the sampling must be dense enough and the spacing of s_i and individual profile points needs to be on the order of the pixel spacing or smaller. The length of the profile is taken greater than the width of the structure to (1) provide sufficient background area to accurately estimate the background values, and (2) account for variations in the width of the structure. The profile extends distances d_1 and d_2, which do not have to be equal, to both sides (Fig. 4.3(a)). Fig. 4.3(b) and 4.3(c) show respectively the sample locations and the resulting patch $\mathbf{P}$.

To determine the right amount to subtract later on, $\mathbf{P}$ is preprocessed to have uniform and zero average background intensity values. The preprocessing is performed first on the whole image and then per structure patch. Global low frequency variations, not associated with $\mathcal{S}$, are eliminated by subtracting a low pass filtered version of the input image. In the case of chest radiographs this removes intensity gradients across the lung that originate from projection of the elliptical shape of the lung in the caudio-cranal direction.

At the patch level the global correction for low frequency variations does not guarantee zero average background values. Therefore a normalization procedure is performed on $\mathbf{P}$ which provides approximate zero intensity values at the borders of the structure patch, i.e. the endpoints of the sampled profiles. A thin plate spline[180] (TPS) surface is fitted through the border points and subtracted from the structure patch. To prevent crossing structures disturbing the TPS plane excessively, border points with outlying intensity values (> 0.95 times the average intensity value of all the border points) were excluded in the fitting process.

Outlier detection

Many of the common decomposition techniques that can be used to reconstruct $\mathcal{S}$ are sensitive to outliers. In our context, outliers are other structures crossing $\mathcal{S}$ corrupting the profiles in $\mathbf{P}$. The presence of a significant number of outliers that dominate over $\mathcal{S}$ leads to inclusion of unwanted sources in $\hat{\mathbf{Z}}_r$ and consequently in $\mathbf{X}_F$. To avoid the effect of these unwanted sources, outlier profiles are detected and remove before performing subspace filtering. Profiles in $\mathbf{P}$ are clustered into two sets $\{\mathbf{P}_1, \mathbf{P}_2\}$ using k-means clustering (Fig. 4.3(d)). As it is impossible for any (unsupervised) method to recover from more than 50% of outliers in a dataset, the set with the largest number of elements, defined as $\mathbf{P}_1$, is assumed to contain the uncorrupted profiles.

Subspace filtering by BSS

A number of techniques have been developed to perform BSS. The most commonly known are Principal Component Analysis (PCA)[181], Singular Value Decomposition (SVD) and Independent Component Analysis (ICA)[182]. PCA and SVD are closely related techniques which give linearly uncorrelated sources. ICA imposes a stronger constraint and produces components which are statistically independent. Non-negative matrix factorization is another technique for BSS where the condition is enforced that the input and the components are non-negative[183]. We assume that the largest variance in $\mathbf{P}$ is caused by $\mathcal{S}$ and therefore use PCA to perform the BSS.

Letting $\mathbf{P}_1$ assume the role of $\mathbf{X}$, the observations from which the sources are computed, Eq. 4.3 is rewritten as

$$\hat{\mathbf{Z}} = \mathbf{P}_1 \mathbf{W}^{-1}. \tag{4.6}$$

PCA determines $\hat{\mathbf{Z}}$ by finding components in the data which are linearly uncorrelated. These principal components are sorted according to their variance. Principal components can be computed by performing an eigenvector decomposition of the covariance matrix $\mathbf{C} = {\mathbf{P}_1}^T \mathbf{P}_1$. Assuming the largest variability in $\mathbf{P}_1$ is caused by $\mathcal{S}$, the first principal components contain this information and $\hat{\mathbf{Z}}_r$ is created by selecting the first n_Z columns of $\hat{\mathbf{Z}}$. n_Z can be set to a fixed quantity or be determined by setting a fixed percentage of the variance f_Z that should be explained by the model.

Filtered profiles can be computed using Eq. 4.4 rewritten as

$$\mathbf{P}_{\mathcal{S}} = \hat{\mathbf{Z}}_r \mathbf{W}, \tag{4.7}$$

where $\mathbf{P}_{\mathcal{S}}$ is a filtered version of $\mathbf{P}$, i.e. an estimate of the projection patch containing only the structure $\mathcal{S}$ (Fig. 4.3(e)). The weights for the profiles in $\mathbf{P}$ are computed by least squares projection

$$\mathbf{W} = \hat{\mathbf{Z}}_r^T \mathbf{P} \tag{4.8}$$

Unrealistically small or large intensity values in profiles can occur in $\mathbf{P}_{\mathcal{S}}$ when elements of $\mathbf{W}$ have large magnitudes. Therefore the weights are constrained by truncation to a fixed absolute maximum magnitude before applying Eq. 4.7,

$$\mathbf{W}_{ij} = \begin{cases} \mathbf{W}_{ij} & \text{if } |\mathbf{W}_{ij}| < \beta \\ \text{sign}(\mathbf{W}_{ij})\beta & \text{if } |\mathbf{W}_{ij}| \geq \beta \end{cases}, \tag{4.9}$$

where β is the maximum value. We refer to β as the PCA model bound, which can be interpreted as the maximum number of standard deviations that each fitted component is allowed to deviate from the mean value.

4.2.2 Suppression

After subspace filtering all profiles in $\mathbf{P}_{\mathcal{S}}$ are assumed to contain mainly intensities originating from $\mathcal{S}$. To remove any noise remaining after subspace filtering the filtered structure patch is smoothed in the s direction using a moving average with a kernel size of σ pixels (Fig. 4.3(f)). The suppression of the structure is then performed at the patch level to create a suppressed patch

$$\mathbf{U} = \mathbf{P} - \mathbf{P}_{\mathcal{S}}^{+}, \tag{4.10}$$

where $\mathbf{P}_{\mathcal{S}}^{+}$ is a positive matrix which is created from $\mathbf{P}_{\mathcal{S}}$ by setting all element values < 0 to 0 (Fig. 4.3(g)). The clipping of the negative values prevents the physically impossible increase of intensity values in $\mathbf{U}$.

$\mathbf{U}$ is projected back into the coordinate system of the original image. In curved structures the sampled profiles typically do not cover all the positions of the whole structure in the original image as the ends of the profiles can be more than one pixel apart in the original image space (Fig. 4.3(b)). This undefined space between the profiles is filled using iterated nearest neighbor interpolation. In this

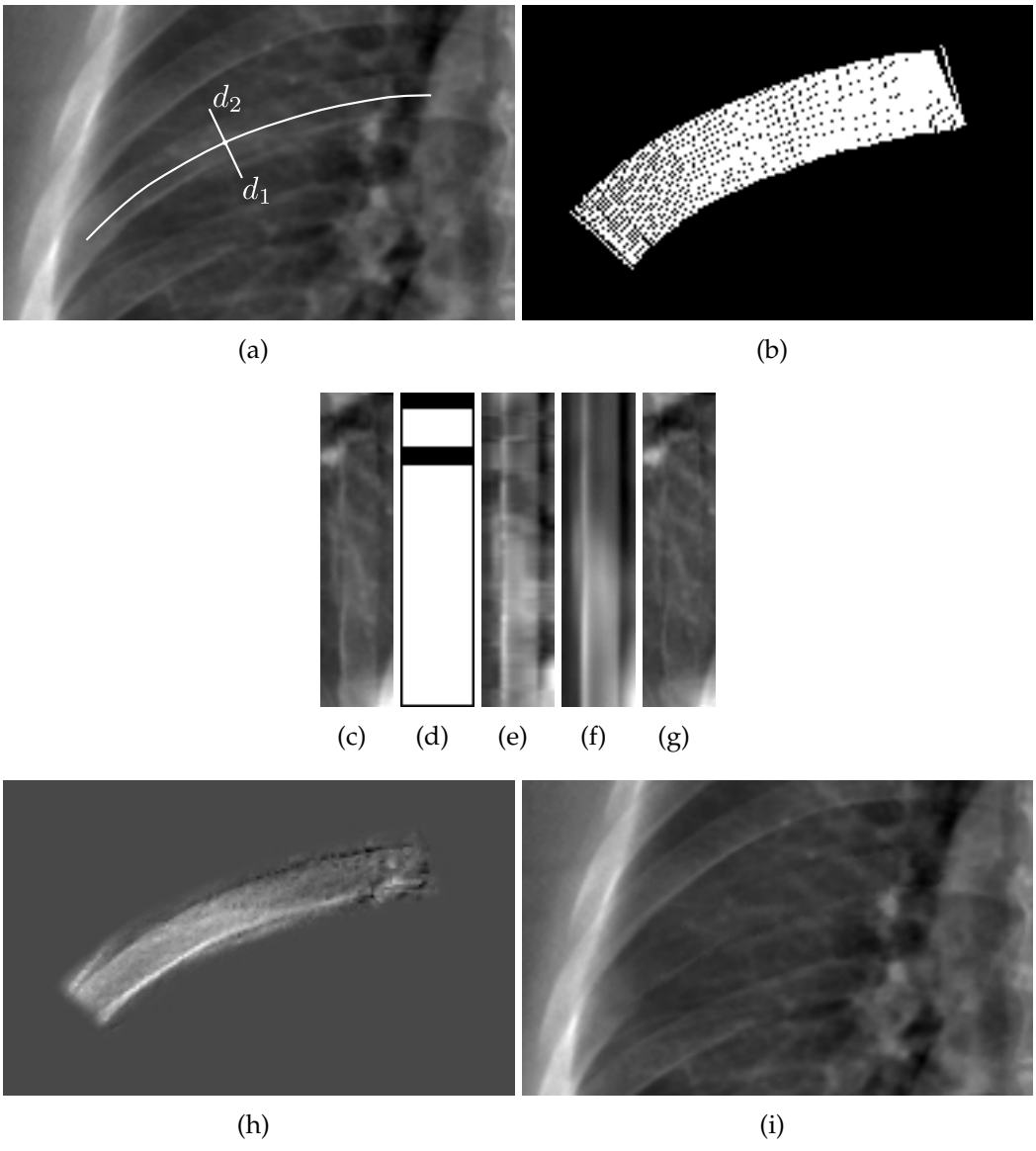


Figure 4.3: Visual overview of method, exemplified by segmentation and suppression of one rib in a simulated chest radiograph. **(a)** Original image with structure $\mathcal{S}$ and aligned elongated curve segment γ. One profile is indicated, extending respectively a distance of d_1 and d_2 to both sides. **(b)** Sample locations of all profiles. **(c)** Sampled profiles arranged in a straightened image patch $\mathbf{P}$. **(d)** Map indicating corrupted profiles (shown in black). Only the uncorrupted white profiles ($\mathbf{P}_1$) are used to construct the PCA model. **(e)** Subspace filtered patch. The filtering is performed using the PCA model. **(f)** Final structure estimation created by applying smoothing to (e) to further reduce noise. **(g)** Result ($\mathbf{U}$) of suppression at the patch level. **(h)** Estimated intensity model of the rib $\hat{L}_{\mathcal{S}}$. **(i)** Final image $\hat{L}_{\mathcal{B}\setminus\mathcal{S}}$ after the suppression of the estimated rib model.

procedure the intensity of undefined pixels, that are 4-connected to defined pixels, are set to the average of their 4-connected defined neighboring pixels. The procedure is repeated until all undefined pixels have been filled.

The intensity values of the suppressed patch in the original image space are used to replace the values in the original image (Fig. 4.3(i)). A fluent transition between the patch and its surroundings is needed to prevent boundary artifacts and is ensured by multiplying it with a Gaussian blurred mask. The mask is created from a binary map indicating the sample locations of the profiles (Fig. 4.3(b)).

4.2.3 Suppression of multiple instances

By repeated application of the algorithm, multiple individual instances of $\mathcal{S}$ can be removed. The resulting suppressed image after removal of the first instance is the input for the removal of the second instance and so on. After removal of all the instances a solution to Eq. 4.2 has been approximated and two estimates of the components of the original image are available: $\hat{L}_{\mathcal{B}\setminus\mathcal{S}}$ containing the background structures and $\hat{L}_{\mathcal{S}} = L - \hat{L}_{\mathcal{B}\setminus\mathcal{S}}$ which contains all the instances of $\mathcal{S}$.

4.3 Experiments & Results

Four sets of experiments were performed to determine the effectiveness of the algorithm and to analyze the effect of parameter changes. In the first experiment ribs are suppressed on chest radiographs simulated from computed tomography (CT) images and compared to simulated rib-free chest radiographs. Additionally the suppression quality was visually evaluated by observers. In the second experiment the effect of suppression is shown on clavicle visibility. In the third experiment the effect of suppression on the visibility of nodules simulated near the clavicles is shown. In the fourth experiment catheters are suppressed in chest radiographs and the quality was visually evaluated by observers.

4.3.1 Rib suppression in chest radiographs simulated from CT

Ribs are the most common projected structure in chest radiographs, and cause a disturbance over the whole image, which makes the detection and analysis of abnormalities and other structures difficult. In this experiment radiographs were simulated from CT images to provide a direct estimate of the suppression quality. The suppression algorithm was run on simulated chest radiographs containing only posterior ribs. The resulting images were compared with simulated rib-free

chest radiographs. The effect of applying subspace filtering, outlier detection and several parameters was evaluated.

Data

From the publicly available ANODE09 database[184] 20 chest CT scans were selected. All scans in the ANODE09 database originate from the NELSON study, the largest CT lung cancer screening trial in Europe. Current and former heavy smokers, mainly men, aged 50-75 years were included in this study. Axial images have a size of 512×512 voxels with a resolution of 0.59-0.83 mm and spacing between axial images 0.7 mm. More details can be found in[61]. The images for this study where selected to contain no gross abnormalities and to be without major rotation of the chest cage.

Segmentation and suppression

Ribs were segmented from chest CT images using the following procedure. Bony structures were selected based on the CT intensity values measured in HU. To prevent artifacts at the transition from bone to other tissue to every voxel a bone probability $p_b(HU)$ was assigned

$$p_b(HU) = \begin{cases} 0 & \text{if } HU \leq 100 \\ (HU - 100)/900 & \text{if } 100 < HU < 1000 \\ 1 & \text{if } HU \geq 1000 \end{cases}$$

An automatic lung segmentation[185] was dilated to encompass the chest cage containing the ribs. The ribs were segmented by selecting voxels with $p_b = 1$ inside the dilated lung mask. This selection will also include parts of the spinal column and the sternum. Individual ribs were segmented by disconnecting them from the sternum and the spine using a manually placed 3D box. Ribs were then divided in their posterior and anterior sections by manually defining a vertical plane running through the widest part of the chest cage.

Simulated chest radiographs were created by an orthogonal projection over the anterior-posterior axis. Only the volume inside the bounding box of the dilated lung mask was projected. Chest radiographs without ribs were created by replacing the segmented ribs with a soft tissue equivalent (HU = 40) in the volume. Partial volume rib voxels ($0 < p_b < 1$) were assigned the same intensity value of 40 HU. Two simulated radiographs per CT case were created: (1) one containing no ribs which was used as reference (Fig. 4.4; second row), and (2) one containing only posterior ribs on which the suppression algorithm was run

Name	Outlier detection	BSS modeling
Full system	yes	yes
No outlier detection	no	yes
Only smoothing	-	no

Table 4.1: Configurations evaluated for the rib suppression experiment.

(Fig. 4.4; first row). After projection, image dimensions were in the range of 346-503 for the x-dimension and 381-498 for the y-dimension.

The segmentation of the ribs in the 2D simulated chest radiograph was performed semi-automatically based on the 3D CT. Individually segmented 3D ribs were projected onto the coronal plane. The centerline of the 2D rib segmentation was determined using the convex sets algorithm described in Staal[186]. The curve segment and the simulated radiograph form the input to the suppression algorithm. Ribs were processed sequentially, i.e. the output image of the suppression of the first rib is used as input for the suppression of the second rib, etc. No particular ordering was present in the segmented ribs. d_1 and d_2 were visually determined from the simulated radiographs and set to 11.25 mm (15 pixels) at both sides, resulting in $M = 22.5$ mm (31 pixels). f_Z was set to 99%. Optimal values for β and σ were determined experimentally (as explained in the evaluation section).

Evaluation

The suppression was evaluated in two subexperiments: (1) by quantifying differences between images before and after suppression, and (2) in a observer experiment.

The amount of suppression was quantified using the sum of squared differences (SSD) between the processed image and the rib-free image by

$$r = \frac{SSD(I_R, I_{NR}) - SSD(I_S, I_{NR})}{SSD(I_R, I_{NR})}, \tag{4.11}$$

where $SSD(.,.)$ is the sum of squared differences between two images, I_R the image with only posterior ribs, I_{NR} the rib-free image, and I_S the result of the suppression algorithm run on I_R. A value of $r = 1$ indicates perfect suppression and $r = 0$ no change. The calculation of r was limited to an area in the upper half of the lung fields with pixels close to the lung border excluded. The reason for excluding this area is that outside this area the segmentation of the ribs fails in

Name	Values	Unit
Smoothing scale σ	16, 32, 48	pixels
PCA model bound β	0.2, 0.5, 1.0, 2.0, 5.0	-

Table 4.2: Parameters evaluated for the rib suppression experiment.

some cases.

Different algorithm configurations, shown in Table 4.1, were evaluated: *Full system* includes subspace filtering and outlier detection, *No outlier detection* system uses subspace filtering but no outlier detection, *Only smoothing* does not perform subspace filtering. Optimal parameter values of the free parameters β and σ for each system were determined using a grid search procedure. The optimization was performed in a leave-one-case out crossvalidation setup where optimal parameters were determined on 19 cases and applied to one case. Table 4.2 shows the tested parameters and their tested values. In total $3 \times 5 = 15$ combination of settings were tested. Differences between optimized configurations were determined by a Wilcoxon signed rank test for the 20 cases.

Additionally, the suppression was evaluated in an observer experiment by three observers: one medical doctor with experience in reading chest radiographs and two certified chest radiograph readers. The observers' ability to discriminate between patches from images simulated without ribs and from images containing ribs was determined before and after suppression. Square patches of 40×40 mm were sampled from inside the unobscured lung fields. Four patches were sampled from the three types of source images derived from 20 cases, giving a total of 240 patches. These patches were presented randomized in one session to each observer who gave a score on the presence of a rib in the patch on a scale of $0 - 100$: 0 and 100 respectively indicating definitely not present and definitely present. Receiver Operating Characteristic (ROC) analysis was performed to determine the observer's performance. The Area Under the ROC curve (AUC) of the two experiment modes was compared using case-based bootstrapping[69].

Results

Fig. 4.4 shows a number of examples of rib suppression on simulated chest radiographs. Visually, most ribs were successfully removed from the simulated radiograph. Table 4.3 compares the amount of suppression r achieved by the tested configurations. *Full system* shows the highest overall improvement and is significantly better than the configurations *Only smoothing* and *No outlier detection*.

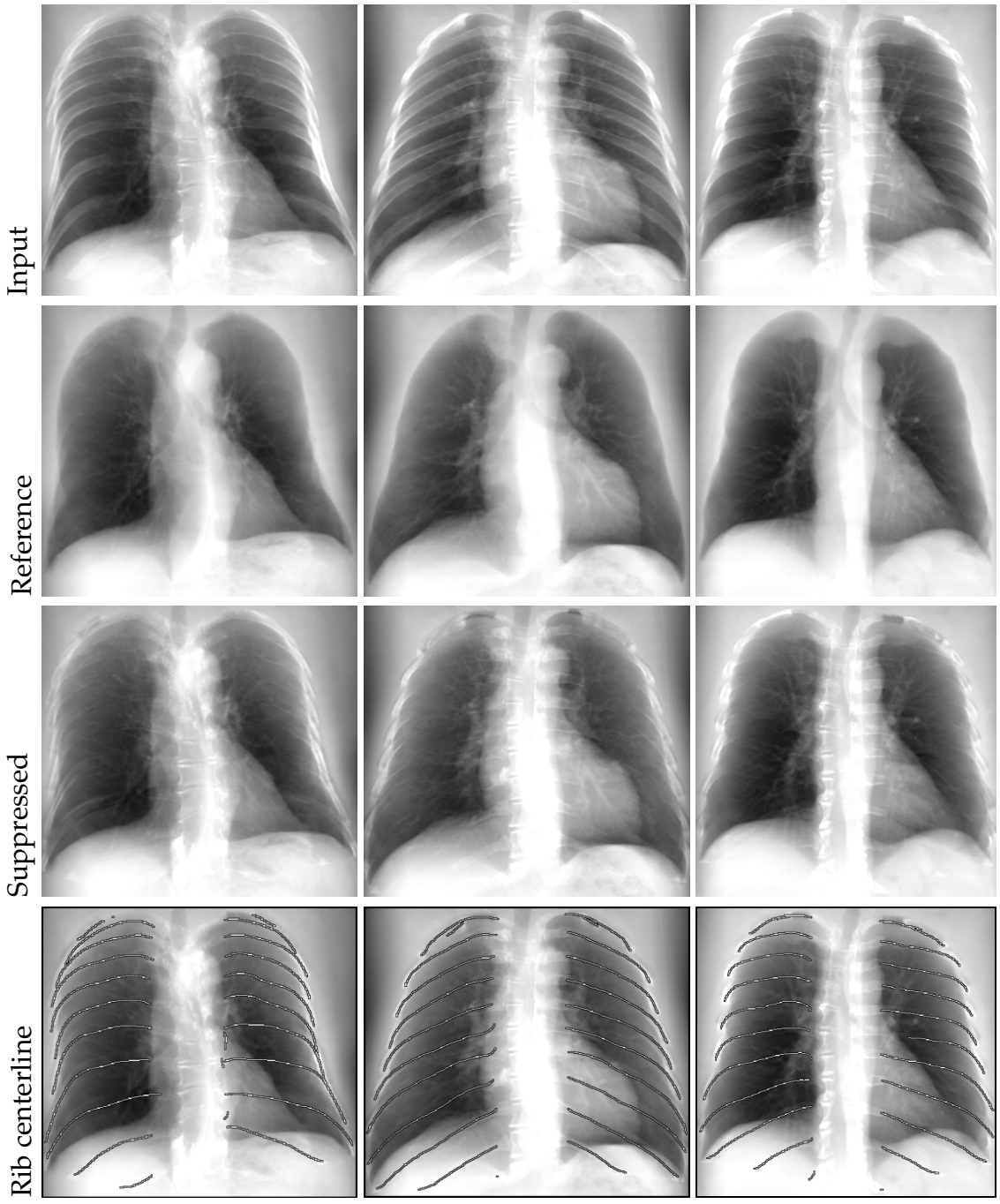


Figure 4.4: Three examples of rib suppression in simulated chest radiographs. On the first row the input of the algorithm, a simulated chest radiograph with only posterior ribs, is shown. The second row shows the reference, a chest radiograph simulated without ribs and other bony structures. The third row shows the rib suppressed image obtained using subspace filtering and outlier detection. The fourth row shows the centerlines that were used as input for the algorithm.

	$r \pm$ std
Full system**	0.851 ± 0.082
No outlier detection*	0.837 ± 0.092
Only smoothing	0.828 ± 0.097

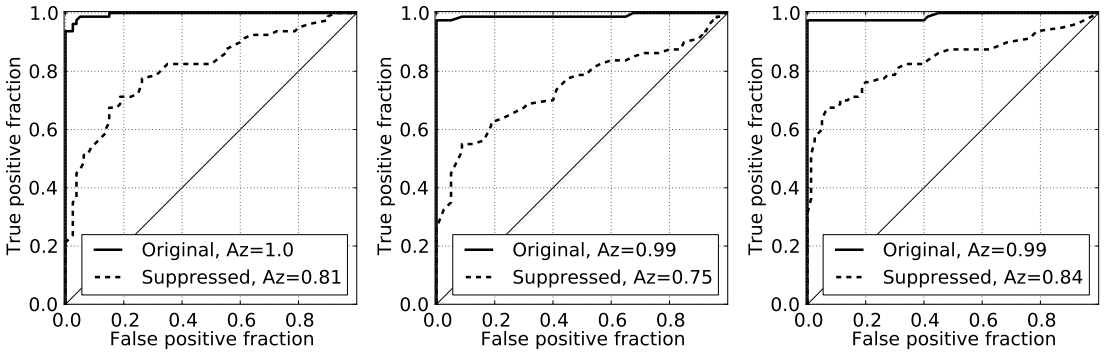
Table 4.3: Performance of the proposed full system for the suppression of ribs and comparison with other system configurations. Systems are ordered according to r, methods which significantly improve over the method below it are indicated with * : $p < 0.05$ or ** : $p < 0.01$. Significance is computed with Wilcoxon signed rank test on the 20 cases.

Name	σ (pixels)	β	Remarks
Full system	32	1.0	$\sigma = 32$ was selected in 12/20 folds
No outlier detection	32	1.0	$\beta = 1.0$ was selected in 18/20 folds
Only smoothing	32	-	Identical in all folds

Table 4.4: Configurations and most selected parameters in crossvalidation for the rib suppression experiment.

No outlier detection performs worse than *Full system* but significantly better than *Only smoothing*. Table 4.4 shows the parameters that were selected the most in crossvalidation for the tested configurations. For subspace filtering $\beta = 1.0$ was most selected in both *Full system* and *No outlier detection*. This value of β limits the model's components to within 1.0 standard deviation of the mean and limits the appearance of crossing structures after filtering. For all three configurations $\sigma = 32$ pixels (± 22 mm) was most selected. This scale is approximately the width of a rib and will remove any remaining small structures, but not smooth away the evolution of the shape of the rib's cross section along the curve segment.

Fig. 4.5 shows the ROC curves for the three observers for judging the presence of ribs in patches extracted from rib free, rib containing, and rib suppressed images. The AUC of the ROC was significantly reduced from very high values on original images to moderate values on suppressed images, respectively from 1.0 to 0.81, 0.99 to 0.75, and 0.99 to 0.84 for observers 1, 2, and 3 with significant differences for all observers (case-based bootstrapping; $p < 0.001$). Before suppression ribs were detected almost without error by the observers. After suppression observers can detect ribs or the remnants in about half of the patches ($< \pm 50\%$; initial steep part of the ROC curve) before starting to confuse patches with and without ribs.



(a) Observer 1 (b) Observer 2 (c) Observer 3

Figure 4.5: Observers' ability to discriminate between rib free patches and patches containing ribs on original and suppressed images. The AUC of the ROC is significantly reduced for all observers comparing original images to suppressed images (case-based bootstrapping; $p < 0.001$).

4.3.2 Suppression of clavicles in chest radiographs

The lung tops are a difficult area to analyze in chest radiographs. Clavicles, ribs, vessels and mediastinal structures overlap and create a complicated pattern in which abnormalities are more difficult to discriminate. The suppression algorithm is used to remove automatically segmented clavicles and evaluated using measures for interior and border conspicuity.

Data

A set of 253 consecutively obtained posterior-anterior chest radiographs were selected from a database containing images acquired at two sites in sub Saharan Africa with a high tuberculosis incidence. The data was previously used to evaluate our clavicle segmentation algorithm[187] (Chapter 3) and is publicly available on `http://crass12.grand-challenge.org`. The data comes from a larger database used for the CAD4TB project, which is aimed at automatically detecting tuberculosis in chest radiographs[188]. All subjects were 15 years or older. Images from digital chest radiography units were used (Delft Imaging Systems, The Netherlands) of varying resolutions, with a typical resolution of 1800 $\times$ 2000 pixels, the pixel size was 250 μm isotropic. The set consisted of both normal and abnormal chest radiographs.

Segmentation and suppression

Clavicle segmentation is performed using the algorithm described in Hogeweg et al.[187]. In this method, supervised pixel classifiers are constructed to segment

the interior, the head and the border of the clavicle. Active shape model segmentation based on the interior segmentation is performed to generate an initial outline. The outline is refined using dynamic programming. The result of the algorithm is an outline with known points corresponding to anatomical positions on the clavicle.

The outline of the clavicle is taken as basis for the suppression algorithm. The outline is divided into three sections: (1) the lower border: running from the edge of the lung field to the start of the head, (2) the head: running from the medial end of the lower border to the medial end of the upper border and (3) the upper border: running from the superior end of the head section to the edge of the lung field. The number of sampled profile points is different on each side of the sections. For the upper and lower border $M = 35$ mm (140 pixels), the profiles are extended $d_1 = 25$ mm and $d_2 = 10$ mm towards respectively the inside and outside of the clavicle making sure that the profiles reach over the other border. For the head section $M = 5$ mm (20 pixels), with respectively $d_1 = 4$ mm and $d_2 = 1$ mm towards the inside and the outside. The other algorithm parameters were set as σ = 11 pixels, f_Z = 99% and β = 0.5.

Evaluation

The visibility of the clavicle before and after suppression was measured with the Weber contrast of the clavicle and with the line response on the border of the clavicles. The two measures reflect the conspicuity of the low frequency interior of the clavicle and the high-frequency borders, respectively. The Weber contrast is defined as

$$C = \frac{I_f - I_b}{I_b} \tag{4.12}$$

where I_f is the average intensity on the clavicle and I_b the average background intensity. I_b is measured in a band around the clavicle with a width of 10 mm. The contrast was measured only inside the unobscured lung fields, which were manually outlined.

The line response is derived from the Hessian matrix[135] calculated at a scale of 0.5 mm. Given the two eigenvalues of the Hessian matrix λ_1, λ_2 with $|\lambda_1| > |\lambda_2|$, the line response is defined as

$$r_l = \begin{cases} 0 & \text{if } \lambda_1 < 0 \\ \sqrt{(\lambda_1^2 - \lambda_2^2)} & \text{if } \lambda_1 \geq 0 \end{cases}, \tag{4.13}$$

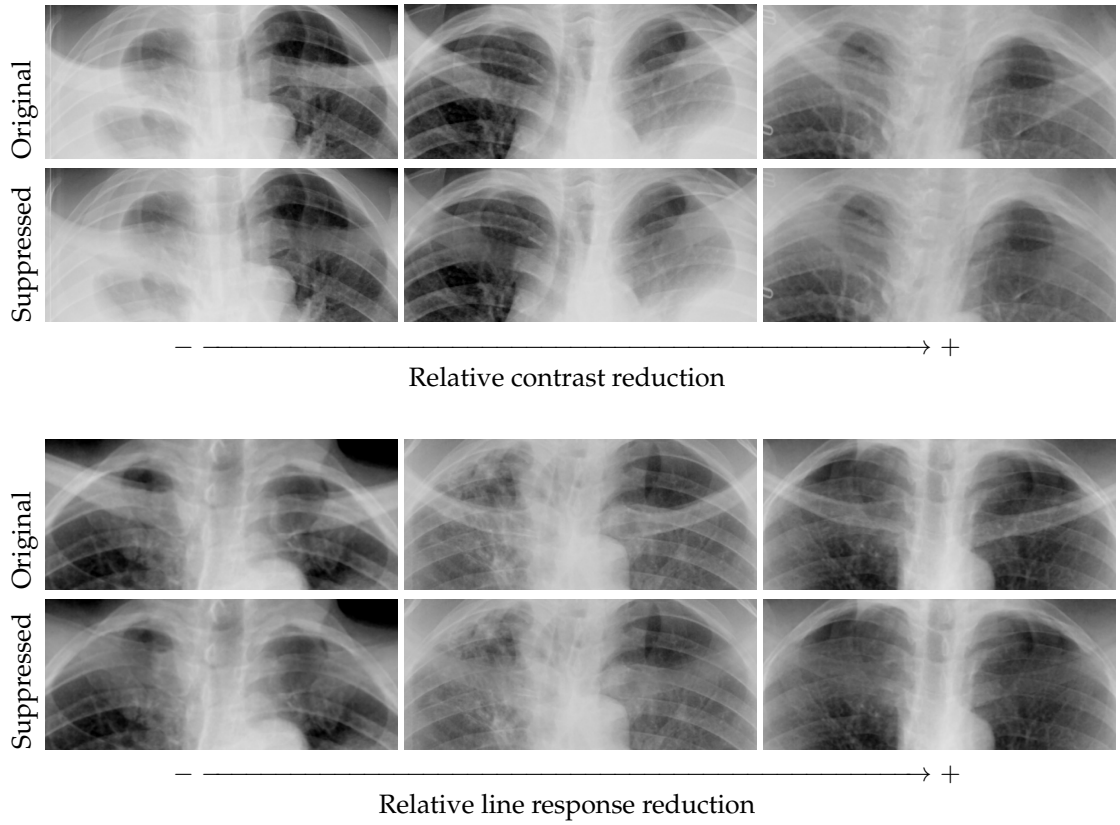


Figure 4.6: Examples of clavicle suppression for six selected cases. The first three cases (top two rows) are sorted on relative clavicle contrast reduction (lowest first). The second three cases (bottom two rows) are sorted on line response reduction (lowest first).

where the condition $\lambda_1 < 0$ ensures that only positive contrast is determined. r_l is measured in a 10 mm wide band centered around the border of the clavicle segmentation and was only computed inside the unobscured lung fields. The measures were computed on original and clavicle suppressed images, with higher values indicating higher conspicuity of the clavicle.

Results

Fig. 4.6 shows examples of original and suppressed clavicles. The interior of the clavicle is mostly suppressed in all cases, while remnants of the clavicle border can still be observed in some cases. The first 3 cases are sorted according to their relative contrast reduction $(C_{\text{org}} - C_{\text{sup}})/C_{\text{org}}$, where C_{org} and C_{sup} are respectively the contrast in the original and suppressed image. Analogously the second 3 cases are sorted according to the relative line response reduction. Over the whole dataset of 253 cases both the contrast of the clavicle body with respect to the

Clavicles					
			Original	Suppressed	
Measure			Mean ± std	Mean ± std	Significance
Weber Contrast	C	$(\cdot 10^{-2})$	2.46 ± 1.11	0.55 ± 0.66	$p < 0.001$
Line response	r_l		1.95 ± 0.57	1.29 ± 0.39	$p < 0.001$

Simulated nodules					
			Original	Suppressed	
Measure			Mean ± std	Mean ± std	Significance
Weber Contrast	C	$(\cdot 10^{-2})$	1.83 ± 0.90	1.35 ± 0.68	$p < 0.001$
Heterogeneity	H	$(\cdot 10^{2})$	1.52 ± 0.51	1.24 ± 0.43	$p < 0.001$

Table 4.5: Change of measures of clavicle and simulated nodule conspicuity from original to suppressed images. Significance is computed with a Wilcoxon signed rank test on 253 and 116 cases for clavicles and nodules, respectively. The line response and heterogeneity measure are only reported for the clavicles and nodules respectively, because they are not relevant for the other structure.

surroundings C and the line response r_l of the clavicle border were reduced significantly (Table 4.5).

4.3.3 Suppression of clavicles in chest radiographs - effect on simulated nodules

Nodules were simulated close to the difficult area around the clavicles and the effect of suppression on nodule conspicuity characteristics was determined. The simulation of nodules enables evaluation on a large set of cases.

Data

Normal chest radiographs were selected from the dataset described in Section 4.3.2 and out of the in total 253 cases 116 contained no abnormalities.

Simulation of nodules

Nodules were simulated in the chest radiograph by projecting CT-derived templates on the clavicle. The procedure was previously described in Snoeren et al.[189] and is summarized here. Five nodules were obtained from a lung cancer screening database[190]. To provide good templates nodules with diameter > 20 mm and which were not connected to the lung wall or large blood vessels were extracted. Nodules were segmented using the smart opening algorithm[191] and extracted as a bounding box. Two-dimensional nodule templates were then created by or-

thogonal ray casting. By scaling and rotation the 3D templates before projection a single 3D template can be used to create multiple 2D templates. To achieve realistic simulation of the nodules a conversion function was used to transform CT units to X-ray units.

Simulated nodules were added to chest radiographs, one per radiograph, with a random nodule template and rotation. The location of the nodule was chosen so that it overlapped with the clavicle, ranging from a slight to full overlap. The contrast of the simulated nodules was set heuristically to a value so that it does not give unrealistically bright nodules but they were still visible for a human observer with knowledge of the location of the nodule. Clavicle suppression was performed with the same settings as in Section 4.3.2.

Evaluation

Nodule contrast is measured similarly as for the clavicles using the Weber contrast (Eq. 4.12), where I_f is the average gray level in the nodule region and I_b in the background region. The simulation of the nodules provides an exact location and direct determination of background and nodule regions. The nodule region is defined as the projected nodule outline. The background region was defined as a 5 mm band around the nodule region. Another aspect of nodule visibility is its heterogeneity. On a uniform background the intensity values of nodules are more homogeneous than when other structures overlap, making them more difficult to detect. The heterogeneity H was defined as the standard deviation of the intensity values in the nodule region.

Results

Fig. 4.7 shows examples of cases with simulated nodules in original images and with suppressed clavicles. It can be observed that the clavicles are substantially suppressed, while the nodules remain visible. Over the whole dataset of 116 cases the contrast of the nodule with respect to the surroundings C was slightly reduced, but the homogeneity of the nodule increased, as indicated by a decrease of heterogeneity (Table 4.5). Before suppression average clavicle contrast was higher than average nodule contrast, but after suppression nodule contrast was more than twice as high. A further analysis of the nodule contrast showed that in the group with an initial high contrast (defined as 50% of cases with highest contrast before suppression) C decreased from $2.04 \cdot 10^{-2}$ to $1.76 \cdot 10^{-2}$ ($p < 0.001$), but in the remaining cases with initial low contrast it slightly increased from $1.09 \cdot 10^{-2}$ to $1.13 \cdot 10^{-2}$ ($p = 0.06$).

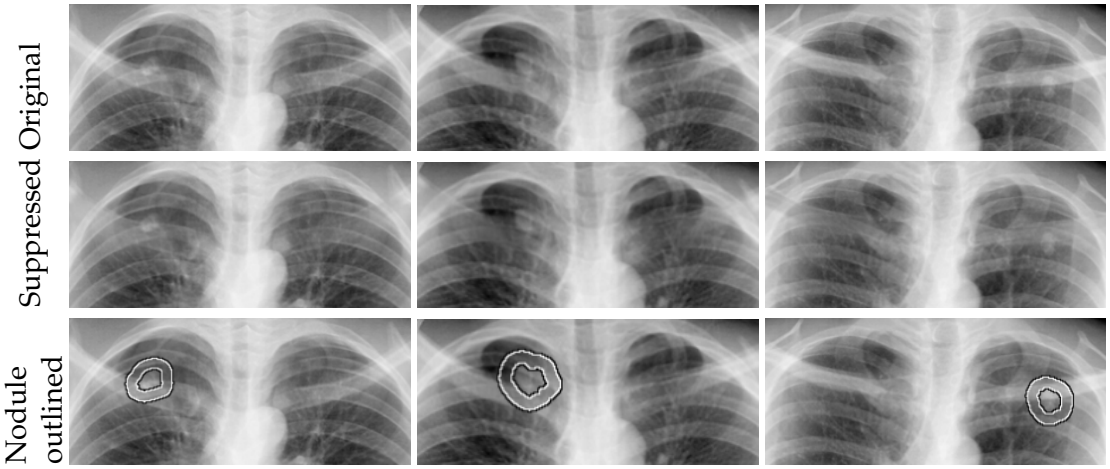


Figure 4.7: Effect of clavicle suppression on nodule contrast for three selected cases. Top row shows the original image with simulated nodule, middle row shows the clavicle suppressed image, bottom row indicates the location of the nodule (inner outline) and the background region for measuring contrast (band between inner and outer outline). The cases are ranked on the relative nodule contrast change from high to low.

4.3.4 Suppression of catheters in chest radiographs

Catheters commonly occur in chest radiographs acquired in a hospital setting and their presence complicates the reading of the images. Catheters were manually segmented and then suppressed using the algorithm. The quality of the suppression was judged by three readers in an observer experiment where they had to discriminate between square patches containing no catheters and patches containing either catheters or containing suppressed catheters.

Data

From the clinical archives of Radboud University Nijmegen Medical Centre, The Netherlands, 36 chest radiographs containing catheters overlapping the unobscured lung fields and 21 chest radiographs without catheters were randomly selected and anonymized. Images were acquired with digital chest radiography units (Siemens Healthcare, The Netherlands) of varying resolutions, with a typical resolution of 2700 $\times$ 2700 pixels and a pixel size of 143 μm isotropic. Suppression was performed on images downsampled to a fixed width of 1024 pixels.

Segmentation and suppression

Catheters were manually indicated by drawing the centerline along its whole length. The centerline was used as input to the suppression algorithm. The settings for the algorithm were determined in a pilot experiment by visual inspection of the suppressed images: σ = 21 pixels, M = 10.5 mm (28 pixels), $d_1 = d_2 = 5.25$ mm, n_Z = 12 and β = 0.5.

Evaluation

The suppression was evaluated in an observer experiment by three observers: one medical doctor with experience in reading chest radiographs and two certified chest radiograph readers. The observers' ability to discriminate between patches with and without catheters was examined in two sessions. In session I patches without catheters and original patches with catheters were presented, in session II patches without catheters and patches with suppressed catheters were shown. Square patches of 30×30 mm were sampled. In chest radiographs containing no catheters patches were randomly sampled from inside the unobscured lung fields. In images containing catheters square patches were sampled along the trajectory of the catheter inside the lung fields, ensuring that the center pixel of the patch coincides with the centerline of the catheter. Three and five patches were sampled from images with and without catheters, respectively, with a total of 213 patches. These patches were presented randomized to each observer who gave a score on the presence of a catheter in the patch on a scale of $0 - 100$: 0 and 100 respectively indicating definitely not present and definitely present. The observers were not aware of the proportion of patches containing catheters in the study. Receiver Operating Characteristic (ROC) analysis was performed to determine the observer's ability to discriminate between patches with and without a catheter. The Area Under the ROC curve (AUC) was compared between session I and II using case-based bootstrapping[69].

Results

Fig. 4.8 shows two examples of catheter suppression inside the lung fields. Visually, the catheter was removed successfully by the suppression over the majority of its length. Fig. 4.9 shows 5 examples of patches used in the observer study. The first 4 patches contained catheters and are sorted on average rating by the three observers. The last example (rating=77) contained no catheter but was rated on average highest on presence of catheters.

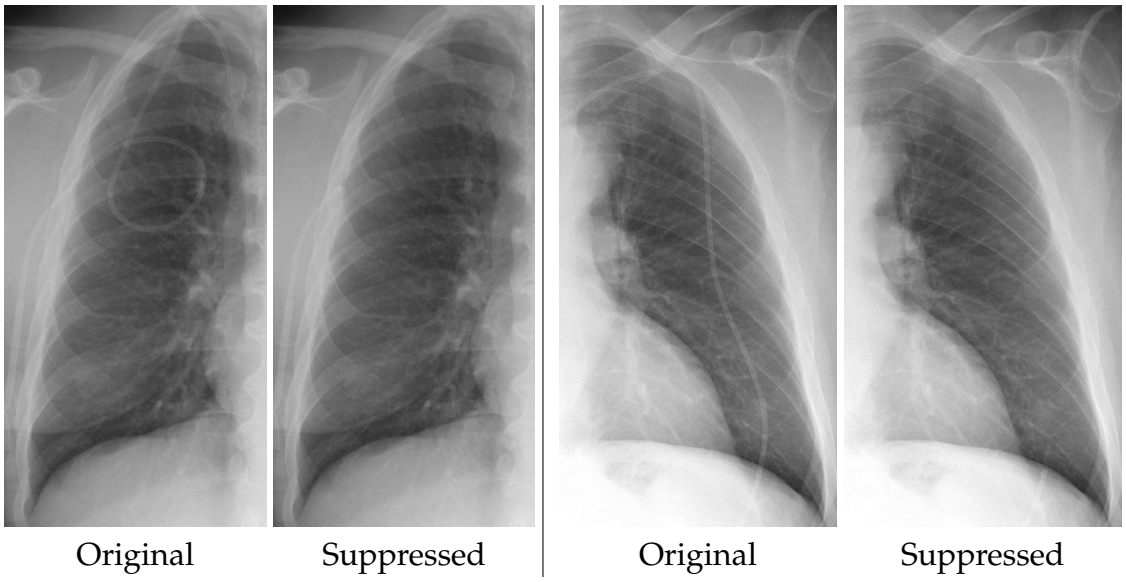


Figure 4.8: Examples of catheter suppression inside the lung fields for two cases. Only the part of the catheter inside the lung fields is suppressed.

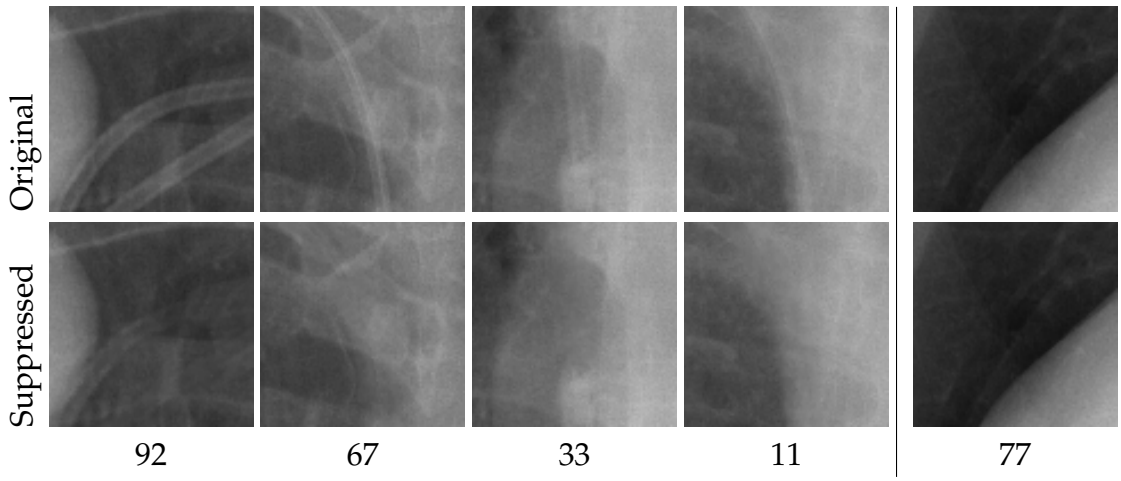
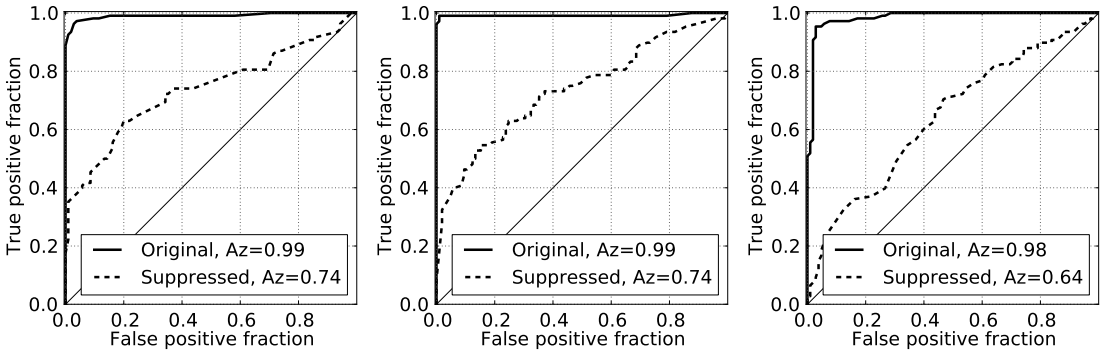


Figure 4.9: Example of four patches with catheters and one without that were used in the observer experiment. The catheter patches are sorted on average score for presence of catheter in the suppressed patch of the three observers. The fifth example does not contain a catheter but shows the highest rated normal patch.



(a) Observer 1 (b) Observer 2 (c) Observer 3

Figure 4.10: Observers' ability to discriminate between unaffected patches and patches containing catheters on original and suppressed images. The AUC of the ROC is significantly reduced for all observers comparing original images to suppressed images (case-based bootstrapping; $p < 0.001$).

Fig. 4.10 shows the ROC curves for the three observers for judging the presence of catheters in patches extracted from catheter free and catheter suppressed images. The AUC of the ROC was significantly reduced from very high values on original images to moderate values on suppressed images, respectively 0.98 to 0.64, 0.99 to 0.74 and 0.99 to 0.74 for observers 1,2, and 3 with significant differences for all observers (case-based bootstrapping; $p < 0.001$). Before suppression catheters were detected almost without error by the observers. After suppression observers can detect catheters or the remnants thereof only in a minority of patches ($< \pm 35\%$; initial steep part of the ROC curve) before starting to confuse patches with and without catheters.

4.4 Discussion

A method to suppress translucent elongated structures in 2D images has been presented. Key elements of the method are subspace filtering of the structure and outlier rejection. The method was evaluated in four experiments on rib, clavicle, and catheter suppression in chest radiographs. In this section we first discuss the results of the four experiments, we subsequentially critically evaluate the merits of the subspace filtering approach and the determination of the background, finally we discuss other applications of the method and consider possible improvements.

In the first experiment it was shown that subspace filtering using PCA improved suppression of ribs in simulated chest radiographs, compared to a method

using only smoothing. The use of simulated radiographs allowed us to measure exactly the amount of suppression, showing a large reduction of the intensity values of the ribs after suppression. In addition an observer experiment was performed were it was found that the ability of human observers to detect ribs in patches after suppression was markedly reduced. Rib centerlines were not automatically segmented in the projected radiograph, but instead derived from the CT segmentation. This allowed us to determine the suppression quality without the influence of errors that typically occur in automatic segmentation techniques. Automatic segmentation of ribs in radiographs has not been fully solved, but a number of systems have been proposed in the literature[78,79,192–195]. Recent other work on suppression of ribs in chest radiographs is based on statistical regression techniques[90,91,196]. In these methods patches with bony structures are replaced with boneless patches by using either massively trained artificial neural networks[90,196] or k-nearest-neighbor[91] regression. Both methods require the availability of dual-energy (DE) bone images as training material. These types of images are not routinely acquired in most settings. Our algorithm does not require the availability of DE images to remove the bony structures. Instead, the information needed to suppress the bone is obtained under only the assumption of the presence of a common profile pattern along the ribs and the clavicles. This property makes the method more easily applicable to other domains where removal of elongated structures is useful.

In the second experiment clavicles were suppressed in chest radiographs. Clavicle suppression reduced clavicle conspicuity as measured by the contrast of the whole clavicle with respect to the background and by the line response on the border of the clavicle. The contrast of the body was reduced to a large extent. The line response gives a measure of the ability of the method to suppress high frequency structures. While the line response is clearly reduced, visual examination shows remaining clavicle borders in some cases. A possible reason for this is that PCA, although in principle able to model any structure up to the Nyquist frequency, has not modeled the border fully. An improvement of the method would be to modify the modeling and the subspace filtering to place extra focus on the border of structures, for example using weighted PCA. For this experiment a fully automatic segmentation was used providing accurate outlines of the clavicles[187] (Chapter 3). Other segmentation methods for the clavicles are available as well[76,77,80,152]. Two other methods to suppress clavicles have been published[80,196]. In Chen et al.[196] a location specific massively trained artificial neural network was

used to suppress the clavicles and good (visually examined) suppression of clavicle body and edges was achieved. As discussed before, a disadvantage of this algorithm is the need for DE images as training material. In Simkó et al.[80] a bone model was created by smoothing along the automatically segmented clavicle border, after which the bone model is subtracted. Such a method, which only uses smoothing to identify the structure, will not be able to cope with larger disturbances: in our first experiment on rib suppression we have shown that outlier detection and subspace filtering, which deal with larger disturbances, significantly contribute to a more accurate suppression.

In the third experiment nodules were simulated in the neighborhood of the clavicles in real chest radiographs. The use of simulated nodules allows for the exact determination of the location of nodules and for creating a larger dataset than would have been possible using public datasets. Nodule contrast decreased slightly after suppression, a similar finding was made by Suzuki et al.[90], who also found a slight degrading of the contrast after suppression. Interestingly, we found that nodules with an initial low contrast showed a slight increase of contrast after suppression. This observation can be explained by the finding that on average the contrast per nodule increased by 15%; the overall slight reduction in absolute contrast is thus mainly caused by large nodules. For detection purposes not only the contrast of the nodule itself, but also the value relative to the overlapping and surrounding structures is important. We found that before suppression the contrast of the clavicle, which is the most conspicuous structure in the upper lung fields, was considerably higher than that of the nodule, but after suppression nodules had on average a higher contrast than the clavicle. Additionally we found that the appearance of the nodule was more homogeneous after suppression. Both the increase of contrast with respect to other structures and the increased homogeneity may aid detection by automatic methods or humans. To achieve these beneficiary effects a perfect visual suppression of bony structures is not required. Instead, the suppressed image provides extra information compared to the original. Providing both original and bone suppressed images to the radiologist is the common mode of operation[85–87] and may help automatic methods as well.

In the fourth experiment human observers judged the quality of the suppression of catheters in chest radiographs. It was found that the observers' ability to identify a catheter was markedly reduced after suppression. In about one third of the patches, originally containing catheters, the observers could still identify

(remnants) of the catheter. Readers rarely take such a close-up look at the radiograph as in this experiment, and the overall suppression of the catheter might be sufficient for practical purposes. The reduction in the observers' ability to detect suppressed structures was higher for the catheters than for the ribs, and this suggests that the proposed algorithm should be extended to improve rib suppression, as is discussed below. Suppression of catheters is a new research area; we are not aware of any previously published method that addresses this topic. In this experiment we used manual segmentations of the catheters, but automatic catheter (tip) detection methods in chest radiographs have been developed[119,197] and could be combined with the presented algorithm to achieve fully automatic catheter removal in chest radiographs. The removal of foreign objects, such as catheters, is also important for automatic processing by computer aided detection algorithms to prevent false positives[198].

In the experiments both real and simulated data were used. In the rib suppression experiment, chest radiographs simulated from CT provided a reference standard which allowed to exactly determine the amount of suppression. Simulation of the clavicles from CT is not possible because the position of the arms in a CT scanner is different from the position in chest radiography, leading to the clavicles being rotated and not overlapping anymore with the lung fields on a posterior-anterior simulation. Instead highly realistic nodules extracted from CT were simulated in the clavicle region to provide an accurate measure of the effect of clavicle suppression on the conspicuity of these lesions, and provide insight into characteristics relevant to their detection in a diagnostic task. As an alternative to simulated data, DE images could be employed. They have as disadvantage that some DE images contain bony structure artifacts as a result of the misaligned subtraction in the imaging procedure[199] and are therefore less suited as reference standard. Another option is to use a digital phantom of known composition, which would allow to exactly measure the amount of suppression achieved. A realistic digital phantom of the chest is difficult to create due to the complexity of the lung structure and to the best of our knowledge none have been described in literature. Instead, a simpler phantom could be constructed, but this makes it difficult to judge the algorithm's merit in a real radiograph.

Subspace filtering, i.e. the use of decomposition techniques to remove the noise subspace of a signal, has been done before using PCA[200], Singular Value Decomposition[201], Independent Component Analysis[178], and non-negative matrix factorization (NMF)[183]. A critical step in these methods is to determine which

and how many components belong to the signal and which to the noise. Under the assumption that most of the variance in the patch originates from the structure of interest, PCA can directly provide the relevant model by selecting the linear components with the highest variance. ICA and NMF do not provide an automatic way to determine the components representing the structure and need an extra step to identify the relevance of each component[178,202].

In certain situations it might be difficult to extract a model which gives a good segmentation of the structure of interest based on only the intensity values in the image patch. An example of such a situation is the presence of many crossing structures, such as ribs crossing another rib, or ribs intersecting the clavicle. In that case a low model dimensionality might not be sufficient to accurately model the structure, as the first few components are used to model the crossing structures. Increasing the dimensionality can partially solve this problem, but will lead to an inclusion of a larger amount of crossing structures in the model. The performance of outlier detection will also be reduced as it will be more difficult to decide which profiles are outliers when their frequency approaches 50% of the dataset. At 50% outlier frequency any (unsupervised) outlier detection technique will reach its break-down point[203]. Potentially this limitation can be remedied by incorporating *a priori* information about the structure of interest, such as by inclusion of a model derived from a larger number of instances. Another situation where the algorithm worked less successfully is when there is a significant change in the appearance of one structure over the course of its centerline. This happens to the part of the ribs close to the chest wall, resulting in a reduced suppression quality. A solution could be to divide the centerline of the structure into multiple segments, so that for each segment a separate model is used.

Estimating the structure of interest through modeling and outlier removal performs significantly better than through smoothing alone, as was shown in the first experiment. We hypothesize that this higher performance was achieved because the noise that disturbs the structure of interest is not purely Gaussian. In a Gaussian noise setting positive and negative disturbances of the structure of interest would cancel out by an appropriate smoothing procedure. High-frequency variations in background tissue density from small vessels and parenchyma can be considered Gaussian and are removed by smoothing. When larger disturbances, such as big vessels and other ribs, cross the structure, smoothing will not cancel out the disturbance but will only distribute it over a larger part of the structure. This is where modeling of the structure provides a better estimation. Larger dis-

turbances cannot be fitted by the model and are not segmented. To ensure limiting the model to the structure of interest, larger disturbances are excluded from modeling by rejecting them as outliers.

A key aspect of a method that suppresses structures by subtraction is to determine accurately the amount of intensity to subtract. This requires the background values to be known. The presented method determines this value from the assumption that the background values have an average of zero. In reality background values are not zero and the patch must first be normalized. The background values are obtained from pixels outside the structure of interest. If a segmentation of the structure of interest is available, the background locations can be easily found. When an accurate segmentation is not available and only a centerline is used as input, the background values are found by sampling profiles that extend well over the expected width of the structure of interest. Care must be taken to not extend so far that other instances of the same structure are included in the patch as this introduces an offset in the background values.

Multiple instances of one type of structure are removed by successive application of the algorithm. The order of removal is not important, provided that in each application only the instance and not other structures are removed. In practice this is not always true and slight differences between different orderings can be observed. An example of this successive application is the suppression of clavicles where first the lower, and later the upper border are used to guide the suppression. This choice was made to focus the suppression on the conspicuous borders.

The proposed method presents a general framework that can be applied to any projection image and to any structure which meets the assumption of translucency and the presence of a common pattern along a curve. It is not necessary for the curve to be located at the center of the structure. As illustrated by the clavicle suppression, the structure can also be decomposed in multiple curve segments to meet the working conditions for the algorithm. Likewise a radial curve can be used to suppress elliptical structures.

The resulting structure suppressed images can be used as an additional image to aid human reading or as input for subsequent processing steps, such as computer aided detection. Clavicle and rib suppression have been shown to improve radiologist's performance in the detection of nodules[85,86]. Rib suppression has been shown to improve the measured visibility of nodules[90], but so far the effect on a fully automatic nodule system has not been determined yet. Another

detection task in chest radiographs where bony structure suppression might be beneficial is identification of infiltrates such as occurring in interstitial disease[147] and tuberculosis[111,188]. A difficult area to interpret in chest radiographs is the hilar region, where the pulmonary vessel tree enters the lung. Vessel suppression might reduce the complexity of the appearance of the lung field and improve abnormality detection in this area. While determining the correct positioning of the catheter in bedside radiographs is a common task in the hospital, its removal might facilitate the automatic screening for, or monitoring of, abnormalities.

4.5 Conclusion

A general method to segment and suppress elongated structures in 2D projection images was presented. The method only requires a curve aligned to the structure of interest and minimal *a priori* knowledge to perform this task. Subspace filtering, based on blind source separation techniques, was used to isolate the structure from the background. The method was evaluated on three tasks – removing ribs, clavicles and catheters in chest radiographs – and showed a marked reduction of the conspicuity of these structures. Experiments with simulated nodules showed an increase of contrast with respect to the clavicles and potential to enhance diagnosis. Future work will focus on additional modeling to further improve suppression.

Acknowledgements

This study was supported by the European and Developing Countries Clinical Trials Partnership (EDCTP), the "Evaluation of multiple novel and emerging technologies for TB diagnosis, in smear-negative and HIV-infected persons, in high burden countries" (TB-NEAT) project.

Quantification of symmetry

5

Abstract

Symmetry is an important feature of human anatomy and the absence of symmetry in medical images can indicate the presence of pathology. Quantification of image symmetry can be used to improve the automatic analysis of medical images.

A method is presented that computes both local and global symmetry in 2D medical images. A symmetry axis is determined to define for each position p in the image a mirrored position p' on the contralateral side of the axis. In the neighborhood of p', an optimally corresponding position p_s is determined by minimizing a cost function d that combines intensity differences in a patch around p and the mirrored patch around p_s and the spatial distance between p' and p_s. The optimal value of d is used as a measure of local symmetry s. The average of all values of s, indicated as S, quantifies global symmetry. Starting from an initial approximation of the symmetry axis, the optimal orientation and position of the axis is determined by greedy minimization of S.

The method was evaluated in three experiments concerning abnormality detection in frontal chest radiographs. In the first experiment, global symmetry S was used to discriminate between 174 normal images and 174 images containing diffuse textural abnormalities from the publicly available CRASS database of tuberculosis suspects. Performance, measured as Area under the Receiver Operating Characteristic curve A_z was 0.838. The second experiment investigated whether adding the local symmetry s as an additional feature to a set of 106 texture features resulted in improvements in classifying local patches in the same image database. We found that A_z increased from 0.878 to 0.891 (p = 0.001). In the third experiment it was shown that the contrast of pulmonary nodules, obtained from the publicly available JSRT database, increased significantly in the local symmetry map compared to the original image.

We conclude that the proposed algorithm for symmetry computation provides informative features which can be used to improve abnormality detection in medical images both at a local and a global level.

5.1 Introduction

Symmetry, a ubiquitous property of both natural and man-made objects, is the property of an object being invariant to certain types of transformations. The most well known forms of symmetry are reflection, or bilateral symmetry, and rotation symmetry. Symmetry as a general feature of objects has been extensively studied in computer vision. Being such a fundamental property of many objects, there have been numerous applications where symmetry has been applied, for example in face detection[204], object tracking[205,206] and analysis of textures[207]. A detailed overview of many aspects of symmetry computation and its applications can be found elsewhere[208]. Many of the proposed algorithms considered symmetrical properties in objects that are described by their boundaries[209–211]. More recently techniques have been developed that detect symmetry directly in images, using point descriptors (features) to measure similarity between symmetric points[212,213].

Although the output of these methods is mainly a binary measure, a continuous symmetry measure is also useful to impose an ordering on a series of objects such that an object with a smaller measure is judged to be less symmetric. A well known example of such a measure is the Continuous Symmetry Measure (CSM) by Zabrodsky et al.[210], which quantifies symmetry in object boundaries. A number of papers have used continuous symmetry measures, such as the CSM, to find correlations with other characteristics. Examples are the relation between facial symmetry and subjective measures of attractiveness[214], fluctuating asymmetry and developmental instability[215], and molecule symmetry and enzyme activity[216].

The human body exhibits a large degree of symmetry, clearly visible on the outside, but numerous organs such as the brain, lungs, and visual system also display symmetry. A loss of symmetry in these organs is often an indication of a disturbance of their normal functioning. For this reason, visual assessment of symmetry in medical images is typically used by human specialists for image interpretation and pathology detection. In automated medical image analysis the use of symmetry has been limited and mainly focused on brain MRI. In Liu et al.[217], symmetry was used to robustly extract the midsagittal plane in pathological brain images. In Sun et al.[218,219], the detection, segmentation, and classification of brain lesions was performed using a symmetry measure that involves computing point-to-point similarities based on the curvature of the gradient vector flow.

Digital subtraction techniques have also been used to show differences between the two sides of the symmetry axis. Li et al.[93,94] performed registration of the left and right lung fields in posterior-anterior chest radiographs, followed by subtraction, to suppress normal symmetrical structures and enhance pathology. In a later study[84], a similar technique was successfully used as a postprocessing stage to reduce the number of false positive detections in a CAD scheme to detect nodules.

Many previously proposed methods rely on the assumption of perfect symmetry. Although many natural objects clearly display properties of symmetry, this symmetry is usually not perfect[220]. In the human body symmetrical organs, such as the brain, are not perfectly symmetrical, even in healthy subjects[221]. The lungs are not fully symmetrical (e.g. the left lung has two lobes and the right one three and the shadow of the heart breaks the symmetry in chest radiographs), but still exhibit a large amount of symmetry in how its internal structures, such as vessels and airways, are organized. The amount of symmetry in a medical image also depends on the properties of the imaging device. Projection radiography, computed tomography (CT), and magnetic resonance imaging (MRI) have different resolutions, contrast, etc. A method analyzing symmetry in medical images should be able to deal with this inherent normal asymmetry.

In this paper, we propose a generic algorithm to assess symmetry in 2D medical images with the aim to detect the presence of pathology. Unlike previous methods, the proposed algorithm deals with both the inherent normal asymmetry of the organs and asymmetry as a result of pathology. The algorithm uses point descriptors and similarity measures to describe the image contents and its symmetry[222–224]. Two symmetry measures are provided: a local symmetry measure, which indicates the level of symmetry in each point of the image; and a global symmetry measure, which summarizes the degree of symmetry of the whole image. These two measures will allow to determine not only the presence of pathology but also provide a localization of the lesion. The potential usefulness of the algorithm is demonstrated in a number of applications involving the detection of abnormalities in chest radiographs.

The paper is organized as follows. Section 5.2 describes the method and its implementation. Section 5.3 describes the experiments and shows results for three different tasks. Section 5.4 discusses the results and in Section 5.5 we conclude.

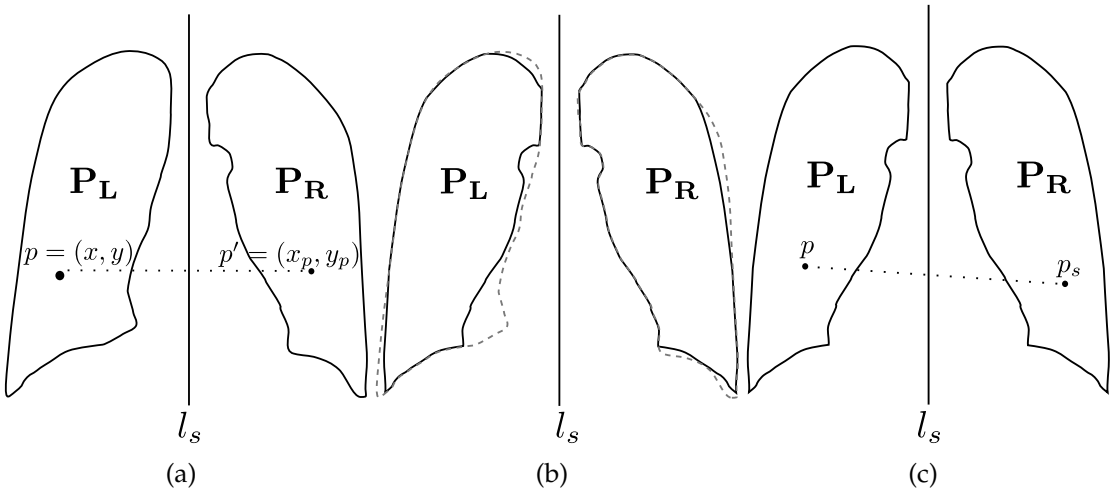


Figure 5.1: Computation of local symmetry $s(p)$ for a point p. (a) Point p in $\mathbf{P_L}$ with mirror symmetric point p' in $\mathbf{P_R}$. (b) Computation is only performed in mirror symmetric sets of locations. The continuous line delimits the points with corresponding mirror symmetric points. The dashed line represents the initial region. (c) The optimal matching point p_s is determined by minimizing a cost function d that measures the dissimilarity of position and image characteristics of p with all points in $\mathbf{P_R}$. The minimum value of d determines $s(p)$. Note that p_s is not necessarily equal to p'.

5.2 Methods

In this section the algorithm to obtain a continuous local symmetry and global symmetry measure is described. Local symmetry s is defined as the minimum dissimilarity between two corresponding points on both sides of the symmetry axis. This is a different definition than the one used in some previous works[213], where local symmetry is described as the presence of symmetrical structures in a subregion of the full image. On the other hand, global symmetry S is defined as the degree of symmetry in the whole image.

5.2.1 Prerequisites

The method operates on a discretized 2D gray value image I in which one mirror symmetric object or region of interest is present. Posterior-anterior chest radiographs and brain MRI or CT sections are common medical examples of such images.

Let l_s be the vertical symmetry axis of the object in I with location $x = x_s$. The image is then divided by l_s into two sets of points: on the left $\mathbf{P_L}$ and on the right $\mathbf{P_R}$ (Fig. 5.1(a)). An equal number of points in both sets is not required and the

correspondences between their points are not known. $\mathbf{P_L}$ and $\mathbf{P_R}$ can cover all the points in the image I (i.e. $\mathbf{P_L} \cup \mathbf{P_R} = I$) or be subsets of points ($\mathbf{P_L} \subset I$ and $\mathbf{P_R} \subset I$, $\mathbf{P_L} \cup \mathbf{P_R} \neq I$), for example presegmented structures of interest such as the lung fields in chest radiographs (Fig. 5.1(a)).

In order to deal with objects which are known *a priori* to exhibit only partial symmetry, we take into account points $p \subset \mathbf{P_L}$ (similarly $p \subset \mathbf{P_R}$) if $p' \subset \mathbf{P_R}$ (similarly $p' \subset \mathbf{P_L}$), where p' is the corresponding point in the reflected position of p with respect to l_s (see Fig. 5.1(b)).

5.2.2 Local symmetry

We define the local symmetry $s(p; l_s)$ of point p with coordinates (x, y) given the symmetry axis l_s as the minimum dissimilarity between p and the points on the contralateral side of l_s:

$$s(p; l_s) = \begin{cases} min\{d(p, p_R), \forall p_R \in \mathbf{P_R}\}, & \text{if } p \in \mathbf{P_L} \\ \\ min\{d(p, p_L), \forall p_L \in \mathbf{P_L}\}, & \text{if } p \in \mathbf{P_R} \end{cases} \tag{5.1}$$

where $d(p_1, p_2)$ is the dissimilarity between points p_1 and p_2. Higher values of s indicate less similarity between points.

Note that the algorithm considers all points on the other side of the symmetry axis to compute $s(p; l_s)$, instead of using only the corresponding point p' located in the reflected position of p with respect to l_s (Fig. 5.1(a)). In this way, the inherent asymmetry in the image is taken into account, in contrast to previously described methods[213,219].

Let $\mathbf{f}(p)$ be a point descriptor which describes the local properties of p and consists of an image component $\mathbf{f_I}(p)$ and a position component $\mathbf{f_p}(p)$:

$$\mathbf{f}(p) = \begin{pmatrix} \mathbf{f_I}(p) \\ \mathbf{f_p}(p) \end{pmatrix}. \tag{5.2}$$

The image component $\mathbf{f_I}$ can be provided by any suitable local descriptor. Examples are SIFT[222], Local Binary patterns[225], etc. In order to determine similar looking positions on both sides of the symmetry axis, computation of $\mathbf{f_I}(p)$ for $p \in \mathbf{P_R}$ is performed on I mirrored locally over the y-axis. In this paper, we define the elements of $\mathbf{f_I}$ as the pixel intensities sampled from a square patch around p with patch size (edge length) m (similar to Avni et al.[224]). The optimal value of m depends on the application and is determined in Section 5.3. In square patches,

and with a row-based sampling, mirroring is performed by reversing the order of elements in $\mathbf{f_I}$.

The position component $\mathbf{f_P}$ consists of the mirror symmetric position of p with respect to l_s:

$$\mathbf{f_p}(p) = \begin{pmatrix} x_p \\ y_p \end{pmatrix}, \tag{5.3}$$

where

$$x_p = \begin{cases} x & \text{if } x \in \mathbf{P_L} \\ 2x_s - x & \text{if } x \in \mathbf{P_R} \end{cases} \tag{5.4}$$

and $y_p = y$. The inclusion of the position component in the descriptor ensures that two points are only considered similar if they have similar visual characteristics and if they are approximately spatially symmetrical.

The dissimilarity $d(p_1, p_2)$ between two points p_1 and p_2 is then defined as the distance K between their descriptors:

$$d(p_1, p_2) = K(\mathbf{f}(p_1), \mathbf{f}(p_2)) \tag{5.5}$$

Many distance definitions are available for K. In this study, the Euclidean distance $K = ||\mathbf{f_1} - \mathbf{f_2}||$ is used.

The elements of the point descriptor contain intensity as well as position values. These values were normalized to zero mean and unit standard deviation over both lung fields before computing distances. In order to control the relative contribution of image and position components, a position weight factor w_p is introduced:

$$\mathbf{f}(p) = \begin{pmatrix} \mathbf{f_I}(p) \\ w_p \mathbf{f_p}(p) \end{pmatrix}. \tag{5.6}$$

High values of w_p favor matching points which have similar symmetric locations. The value of w_p depends on the application and optimal values are studied in Section 5.3.

Determining the local symmetry using Eq. 5.1 can be computationally expensive if the number of points is large. To reduce computational requirements, finding the most similar point is formulated as a 1-nearest-neighbor problem. Efficient solutions to this problem have been developed which precompute data structures and provide a fast approximation close to the exact solution[226]. In this

work we use the Approximate Nearest Neighbor (ANN) algorithm described in Arya and Mount[136]. The algorithm uses precomputed kd-trees and provides an approximate solution which insures that the distance to the approximated nearest neighbor is smaller than $(1+\epsilon)$ times the distance to the true neighbor. In this work $\epsilon = 2.0$ is used.

5.2.3 Global symmetry

The global symmetry measure $S(I; l_s)$ of an image I given the symmetry axis l_s and the sets $\mathbf{P_L}$ and $\mathbf{P_R}$ is computed by averaging all the local symmetry measures $s(p; l_s)$ in $\mathbf{P_L}$ and $\mathbf{P_R}$. If the set of all N locations on both sides of the symmetry axis is defined as $\mathbf{P} = \mathbf{P_L} \cup \mathbf{P_R}$, S is then defined as:

$$S(I, l_s) = \frac{1}{N} \sum_{p \in \mathbf{P}} s(p; l_s) \tag{5.7}$$

Low values of S indicate overall similarity of image characteristics on both sides of the symmetry axis. A value of $S = 0$ indicates that for every point a perfect analog has been found on the other side at the expected reflected position. High values of S indicate the presence of image characteristics on one side which cannot be found on the other side.

Local symmetry values s are spatially correlated and contain redundant information. The global symmetry can therefore be estimated using only a subset of $\mathbf{P}$ without losing its discriminative properties. In Section 5.3 we determine the effect of reducing the number of locations used in the computation of S by sampling every $\sqrt{\kappa}$th pixel in the x- and y-direction; thus subsampling $\mathbf{P}$ with a factor κ.

5.2.4 Determination of optimal symmetry axis

In medical imaging, the scanning protocol typically ensures that anatomical structures have a fixed orientation and location in the image. For example, in posterior-anterior chest radiographs the lung fields are centered and the caudo-cranial direction of the patient is aligned with the y-dimension of the image. For brain imaging with computed tomography or magnetic resonance imaging a similar fixed relation between patient and image coordinate systems is common. In practice, locations of axes or planes of symmetry are not exactly known and are not necessarily aligned with the image axes.

We estimate the optimal position of the symmetry axis l_s from an initial ap-

proximation by minimization of the global symmetry value S. Note that the ensuing discussion relates to 2D images, but the procedure can be easily extended to higher dimensions.

In-plane rotation

If the patient is rotated in the xy plane, it will cause the symmetry axis to deviate from the verticality which is expected by the algorithm. In order to identify the optimal angle of the symmetry axis, anatomical structures are rotated upright by artificially imposing a range of rotations with different angles to the image. The angle which results in the minimum global symmetry S corresponds to the upright position. Let α be the angle used to rotate the image I in the xy plane around the image center. Let $S(I; l_s, \alpha)$ denote the global symmetry value computed for I after rotation. The optimal angle α_{opt} is defined as:

$$\alpha_{\text{opt}} = \underset{\alpha \in A}{\operatorname{argmin}}\, S(I; l_s, \alpha) \tag{5.8}$$

where A is a set of test angles. Fig. 5.2 shows an example of in-plane rotation for a chest radiograph.

Symmetry axis x-coordinate

The x-position of the initial approximation of the symmetry axis x_s (Sect. 5.2.5) can be displaced from its optimal position. A similar minimization procedure as for the in-plane rotation was used to find the optimal position. Let $S(I; l_s, \delta)$ denote the global symmetry value of image I after applying an horizontal displacement δ to the symmetry axis l_s. The optimal horizontal displacement δ_{opt} of l_s is computed as

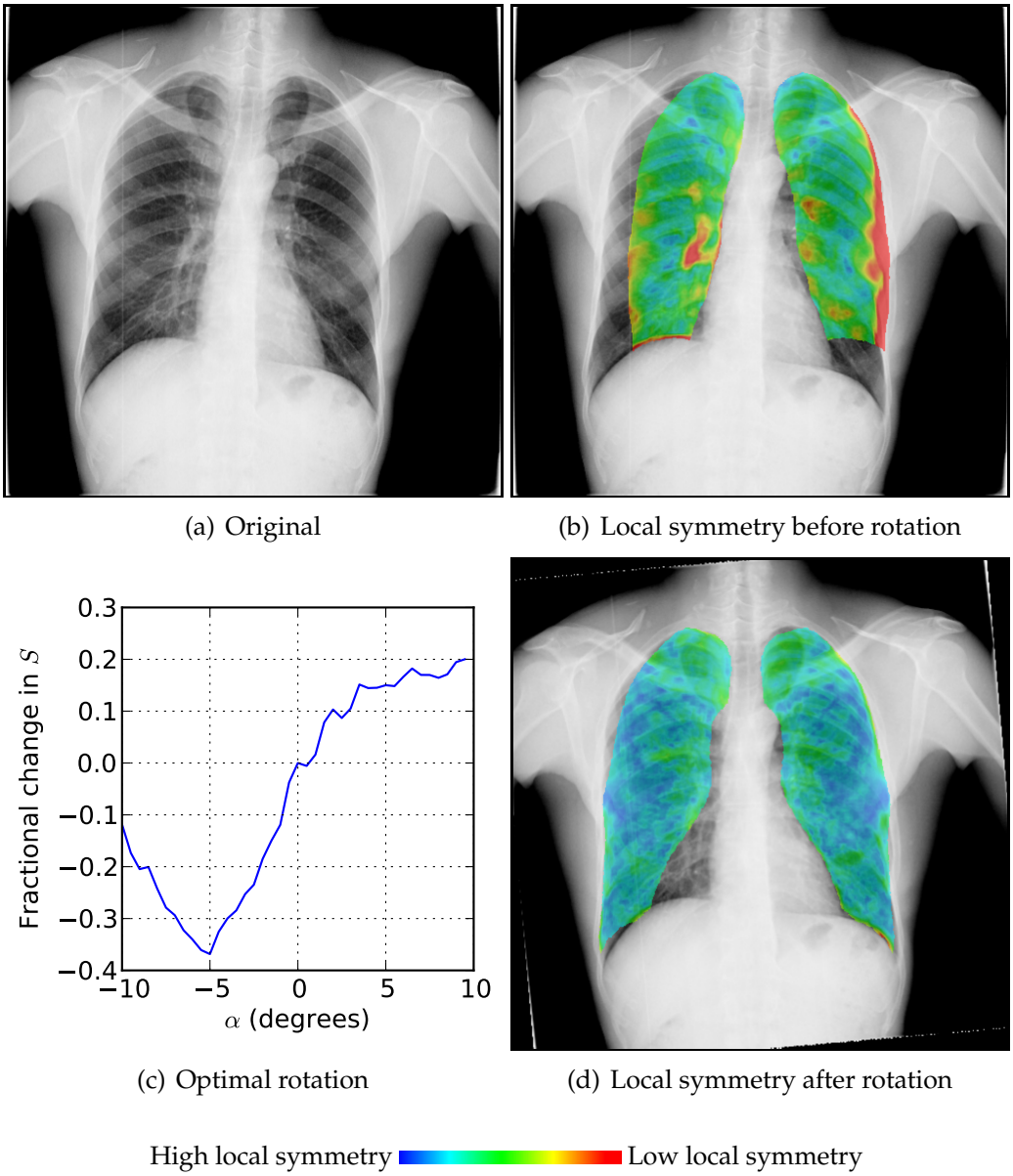
$$\delta_{\text{opt}} = \underset{\delta \in \Delta}{\operatorname{argmin}}\, S(I; l_s, \delta) \tag{5.9}$$

where Δ is a set of test locations.

5.2.5 Symmetry computation in chest radiographs

In this paper, we select the analysis of chest radiographs to evaluate the performance of the proposed symmetry measures in real medical images. Specific details for symmetry computation in chest radiographs, which are used in the experiments, are given in this section.

The expected scale of normal and abnormal structures determine the working resolution of the images for symmetry computation and the scale at which



(a) Original (b) Local symmetry before rotation

(c) Optimal rotation (d) Local symmetry after rotation

Figure 5.2: Example of local symmetry maps before and after optimal rotation in the xy plane. Color maps in (b,d) indicate local symmetry. Color scale has arbitrary units, because s values have only a relative interpretation. The graph in (c) indicates the relation between the rotation angle α and S (compared to the original image). The image rotated by the angle corresponding to the minimum value of S in (c) is used in (d).

$\mathbf{f_I}$ is computed. Chest radiographs were resampled to a fixed width of 512 pixels in all experiments. Optionally, images can be preprocessed to enhance certain structures. The use of a normalization procedure and its properties may influence the symmetry measures. In Section 5.3 we evaluate the use of a local normalization procedure which enhances contrast and removes low frequency variations[132]. This procedure locally normalizes the intensity deviation from the average to the local standard deviation:

$$I_{LN} = \frac{I - \widetilde{I}}{\sqrt{\widetilde{I^2} - (\widetilde{I})^2}} \tag{5.10}$$

where I indicates the original image, I_{LN} the locally normalized image and $\widetilde{(\cdot)}$ blurring by convolution with a Gaussian kernel with scale σ_N.

An initial location of the vertical symmetry axis is determined as follows. The existence of a binary segmentation of the lung fields is assumed, where lung fields have value 1 otherwise 0. A one-dimensional projection image is created by orthogonal averaging over the y-direction. Interpreted as an 1D function this image contains two maxima, corresponding to the lung fields, and three minima corresponding to the two parts at the sides of the image and the part between the lung fields. The approximate position of the symmetry axis is determined by the minimum value in the middle 20% of the curve which corresponds to the part between the two lung fields.

Computation of the symmetry measures was performed using locations in the symmetric lung fields only (as in Fig. 5.1(b)). Please note that symmetric locations are redetermined in each iteration of the symmetry axis optimization.

5.3 Experiments & Results

In this section, we describe the results of three experiments for the detection of pathology in chest radiographs using global and local symmetry measures. In the first experiment global symmetry was used to discriminate between normal and abnormal images. In this experiment the influence of algorithm parameters on the final result was extensively studied. In the second experiment the contribution of the local symmetry measure to a set of general texture features was determined in an image patch classification task. In the third experiment the local symmetry measure was used to enhance nodule contrast.

5.3.1 Global symmetry to discriminate between normal and abnormal images

The global symmetry measure S quantifies the presence of symmetrical structures on both sides of the symmetry axis. Chest radiographs (CXRs) of healthy persons are largely symmetric and tend to give low values. The presence of abnormalities in the lung fields tends to increase S. In this experiment, we study the discriminatory power of S to distinguish between normal and abnormal images.

Data

A set of 348 CXRs (174 normal, 174 containing textural abnormalities) was selected from a database consisting of images of tuberculosis (TB) suspects. Images from digital chest radiography units were used (Delft Imaging Systems, The Netherlands) of varying resolutions, with a typical resolution of 1800 $\times$ 2000 pixels, the pixel size was 250 μm isotropic. The set is a subset of a publicly available database described in Hogeweg et al.[187] where normal images and images containing textural abnormalities inside the lung field were selected. The normal/abnormal decision is based on the absence or presence of textural abnormalities in the image (see Section 5.3.2).

A lung segmentation are used to determine the initial symmetry axis and limit the computation of local symmetry. They were obtained using a previously developed algorithm, which is a combination of pixel classification and shape modeling[76].

Experiment

The discriminatory power of S to distinguish normal and abnormal images was evaluated using the area A_z under the Receiver Operating Characteristic (ROC). In this experiment the influence of the algorithm parameters, namely patch size m, position weight w_p and subsample factor κ was studied. The effect of varying parameter values was first determined per individual parameter; as starting values for each experiment we used $w_p = 10.0$, $m = 9$ pixels, $\kappa = 1$. These values were found to work well in a patch-based categorization and retrieval method involving chest radiographs by Avni et al.[224]. After this first approximation the optimal combination of w_p and m was determined. The set of angles A used to determine α_{opt} was $\{-10.0, -9.5, \ldots, 9.5, 10.0\}$ degrees. The set of horizontal displacements Δ used to estimate δ_{opt} was $\{-10, -8, \ldots, 8, 10\}$ pixels.

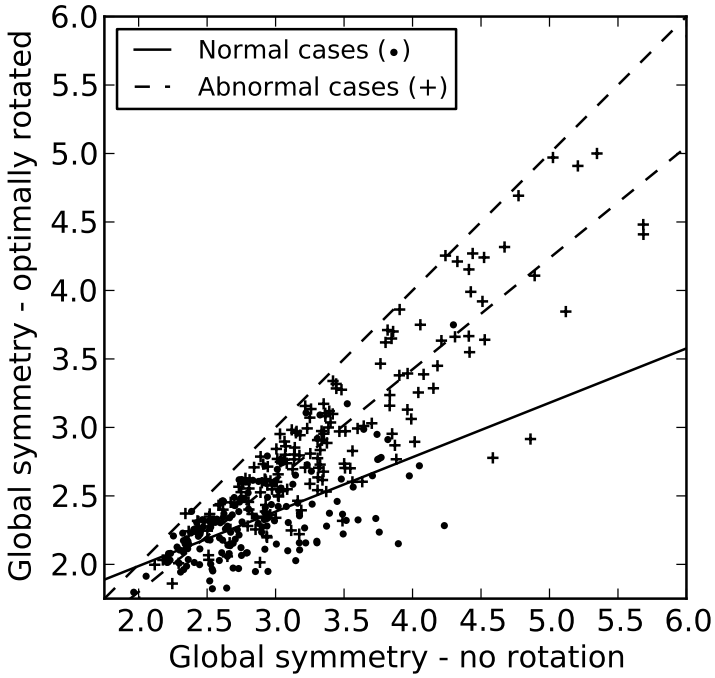


Figure 5.3: Effect of optimal rotation on global symmetry scores S for 348 cases. Scores were computed with optimal parameter values. S for normal cases (dots) decreases relatively more than abnormal cases (pluses) as shown by the trendlines. As a result, projection on the horizontal axis (no rotation) results in high overlap of normal and abnormal cases, but projection on the vertical axis (optimal rotation) yields good separation.

Results

Fig. 5.3 shows the effect of optimal rotation on individual cases for the default parameter settings. By definition all scores are equal or lower after optimization. For normal cases (green) the scores are reduced more than for the abnormal cases (red), as indicated by their respective trendlines. This differential change between normal and abnormal cases leads to a large improvement in discriminative performance.

Fig. 5.4 shows the effect of varying the free parameters on A_z for the 348 test images. Optimal rotation of the symmetry axis was important for the majority of parameter values and led to a large increase in performance. Additional optimization of the x-location did not lead to large further increases. Performances reported in the remainder of this section refer to the images with optimal rotation and x-location. A_z showed a slow increase with increasing m up to a value of 13 pixels and followed by a slow decrease. In the m range of 9-21 pixels, A_z values were highly stable. At $w_p = 5.62$ the optimal A_z of 0.835 was achieved. In the

w_p range of 3.2-32 A_z was mostly stable with values > 0.82. Higher and lower values of w_p led to a reduction in performance. Especially for low values of w_p, in which case the influence of $\mathbf{f_I}$ increases relatively to $\mathbf{f_p}$, performance decreased substantially. The relation between κ and A_z was mostly stable for $\kappa \leq 16$, for $\kappa > 16$ performance was slightly reduced until it breaks down at $\kappa = 256$.

The value of w_p is related to the value m via Eq. 5.6; higher values of m require higher values of w_p to maintain the same weighting of intensity and position information. To reflect this, all combinations of $m = (9, 11, 13, 15, 17)$ and $w_p = (1.77, 3.16, 5.62, 10.0, 17.7, 31.6)$ were tested with $\kappa = 16$ to determine the optimal combination. For these parameter values the highest performance was found in their individual optimization. The optimal combination was $(m, w_p) = (15, 17.7)$ with $A_z = 0.838$; these parameter values were used in subsequent experiments.

Computation times are related to the amount of subsampling. Computation times (at a single core of a Core 2 Duo @ 3.0 Ghz) decreased from 50 s ($\kappa = 1$) to 13 s ($\kappa = 4$) and 8 s ($\kappa = 16$) for an average case ($\pm 40,000$ positions), using optimization of image rotation and x_s, and with optimal (m, w_p) values. Subsampling with a factor $\kappa = 16$ gives minimal performance loss compared to $\kappa = 1$. An additional reduction of computation time can be achieved by not performing x_s optimization, which had only a small effect on performance.

5.3.2 Local symmetry in combination with texture analysis

In this experiment the effect of adding local symmetry to a set of texture features when classifying patches and images for the presence of textural abnormalities was evaluated.

Data

The dataset is the same as used for the first experiment (Section 5.3.1). For training the patch classifier labeled examples of patches are required. Manual outlines of abnormalities were created in the full set. Patches whose center is inside the outline were assigned the label abnormal. Normal patches were only sampled from normal images. From the original images 145,315 patches (116,252 normal and 29,063 abnormal) and from the optimally rotated images 144,905 (115,924 normal and 28,981 abnormal) patches were extracted, both with a normal to abnormal ratio of exactly 4:1.

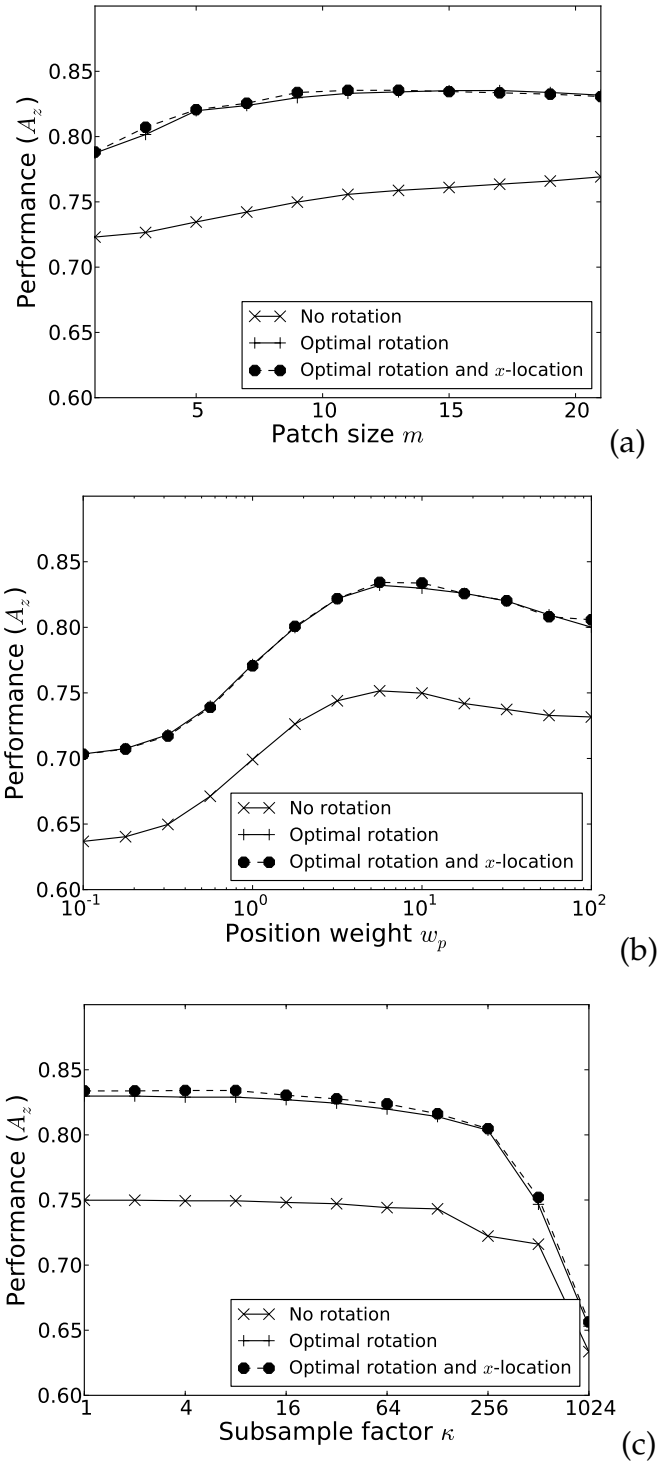


Figure 5.4: Optimization of position weight w_p, subsample factor κ, and patch size m for global symmetry computation on a set of 348 images. Image classification performance for original images, optimally rotated images, and optimal x-location of the symmetry axis are shown as function of the three parameters: (a) Patch size (b) Position weight (NB the x-axis is logarithmically scaled). (c) Subsample factor (NB the x-axis is logarithmically scaled).

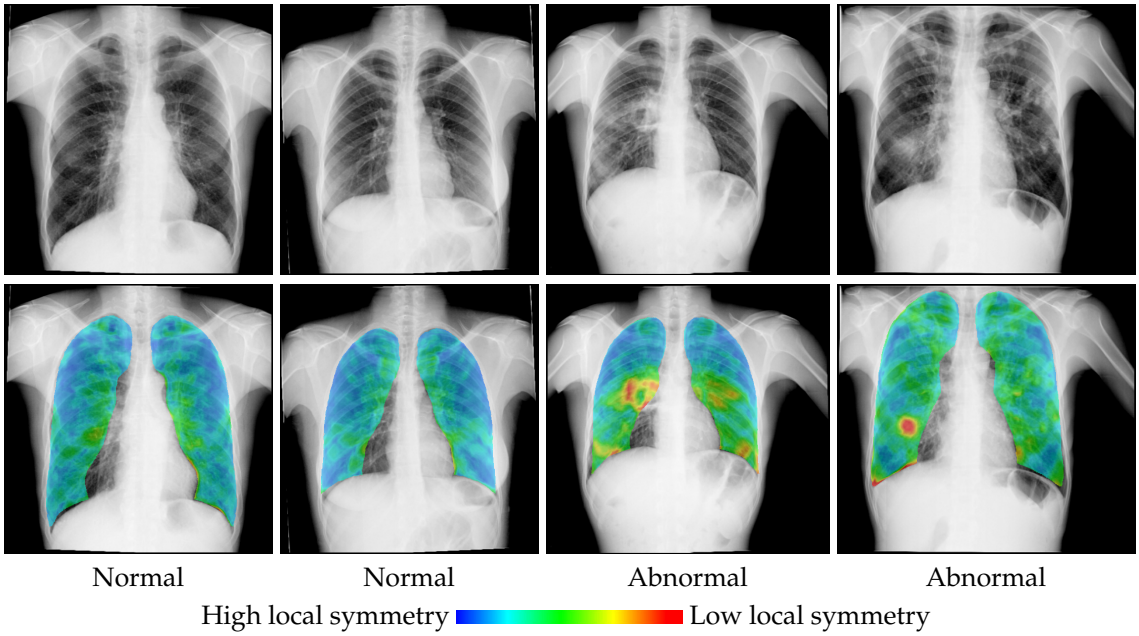


Figure 5.5: Examples of local symmetry for two normal and two abnormal images. Color maps indicate local symmetry; color scale has arbitrary units, because s values have only a relative interpretation. The scaling is identical in all the examples.

Experiment

Local symmetry maps were computed with the optimal parameter values determined in Section 5.3.1. The image rotation and symmetry axis x-location were optimized. The detection of textural abnormalities is based on texture analysis of circular image patches (radius $= 32$ pixels) sampled every 4 pixels. Texture features were computed by extracting statistics of Gaussian derivative filtered images of order 0 through 2 (L, L_x, L_y, L_{xx}, L_{xy}, L_{yy}), at scales 1, 2, 4, and 8 pixels. The first four moments (mean, standard deviation, skew, and kurtosis) of the intensity distribution of each Gaussian derivative filtered image and the original image were computed from pixels inside the corresponding circular patch. This method has recently successfully been used to detect textural abnormalities related to TB in chest radiographs[112,148]. Two general position features, namely the x- and y-position normalized to the image size, and four lung segmentation derived position features, namely the x- and y-position normalized to the bounding box of the lung fields, the distance to the lung boundary and the distance to the center of gravity of the lung fields, were added to the texture features. A total of 106 features per patch were extracted[148]. Image patches were sampled inside the segmented lung fields and assigned an abnormality likelihood with a Gen-

tleBoost classifier[58] which used 100 regression stumps as weak classifiers. Image locations outside the mirror symmetric lung fields were assigned $s = 0$. Images were assigned an overall texture score by computing the 95th percentile of the cumulative distribution of patch likelihood scores[148]. This texture score was used to determine image classification performance.

Two sets of features were compared: (1) the texture+position features totaling 106 features, and (2) the texture+position features and local symmetry totaling 107 features. These feature sets were compared in a patch classification and image classification experiment. In addition, the performance of local symmetry as a single feature was determined in the patch classification experiment. Classification performance was determined using A_z. Significant differences were determined with case-based bootstrapping[69] using 1,000 bootstrap samples. Training and testing of the 348 cases was performed in 2-fold crossvalidation.

Results

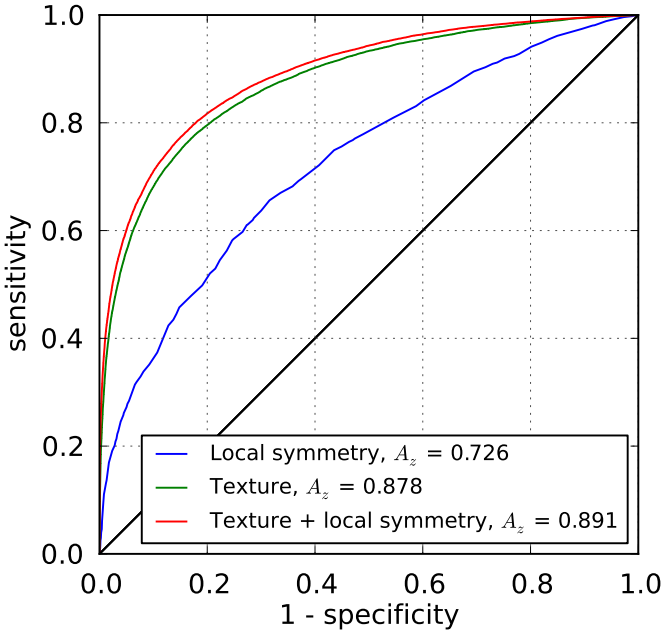
Figure 5.5 shows four examples of CXRs: the first two contain no textural abnormalities, the last two contain several abnormalities across the lung fields. Local symmetry maps are shown for all four cases. In the normal cases values of s are on average low, with slightly higher values close to the hilar structures. In the abnormal cases abnormalities are highlighted in the local symmetry map.

Figure 5.6(a) shows the results of the patch classification experiment. Local symmetry as a single feature achieved $A_z = 0.726$. The texture+position features achieved $A_z = 0.878$. The addition of the local symmetry features to the texture+position features significantly increased performance to $A_z = 0.891$ ($p = 0.001$).

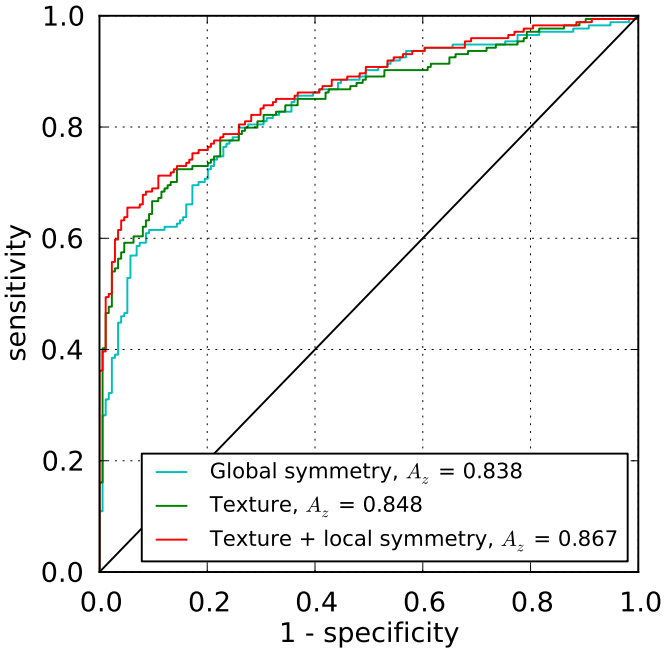
Figure 5.6(b) shows the results of the image classification experiment. The texture+position features achieved $A_z = 0.848$. The addition of the local symmetry features to the texture+position features significantly increased performance to $A_z = 0.867$ (p = 0.01). For comparison the results of using global symmetry to classify images is also shown in the figure. No significant difference was found in performance between global symmetry and the texture+position features alone ($p = 0.352$).

5.3.3 Local symmetry to detect nodules

Lung cancer is commonly detected on radiographs, but it is known that retrospectively visible lesions are missed by radiologists in 19-90% of cases[151], for this



(a) Patch level performance



(b) Image level performance

Figure 5.6: Effect of adding local symmetry to a supervised system detecting textural abnormalities analyzed using ROC analysis. Experiments were done on 348 chest radiographs. *Texture* is the basic system without local symmetry, *Texture + local symmetry* includes s as a patch feature. (a) Patch (local) level performance. *Local symmetry* is the system with s as the only patch feature. (b) Image level performance. *Global symmetry* is added for reference and shows the performance of S as a single image feature.

reason computerized support for lung nodule detection is an active area of research[97]. Some nodules are very difficult to detect, both by computer and by human readers, and a careful comparison of the left and right lung fields is often required to discern them. An experiment was performed to determine changes in contrast of nodules on CXRs after symmetry computation. In this experiment also the effect of preprocessing with local normalization was determined. Local normalization improves contrast of nodules[132], and also serves to reduce low frequency variations which are uninformative for the detection of small lesions.

Data

For evaluation we used the publicly available JSRT database consisting of 93 normal cases and 154 abnormal cases[37]. Only abnormal cases were used in this experiment; each contained one nodule of which location and radius were available. Images were digitized 12-bit posterior-anterior CXRs, scanned to a resolution of 2048×2048 pixels with an isotropic pixel resolution of 175 μm. Nodule sizes ranged from 5 to 60 mm (median = 15 mm) with varying degrees of conspicuousness. Four cases, in which the nodule was located outside the lung fields, could not be used; thus we had 150 cases available for analysis.

Experiment

Images were resampled to a width of 512 pixels. Local symmetry maps were computed with the optimal parameter values determined in Section 5.3.1. The image rotation and symmetry axis x-location were optimized.

The visibility of a nodule was determined by its contrast with its neighboring background, using the Weber contrast

$$C = \frac{I_f - I_b}{I_b} \tag{5.11}$$

where I_f is the average intensity of the nodule region of interest (ROI) and I_b the average background intensity. The nodule ROI is defined by a circle centered at the nodule location and with a radius r obtained from the JSRT annotations. I_b is measured in a band enclosing the nodule ROI with a width of $0.5r$. The contrast was measured only inside the unobscured lung fields, which were automatically segmented using active shape models[76].

C was computed on four types of input images: the original image, the locally normalized (LN) image with $\sigma_{LN} = 16$ pixels[132], and local symmetry maps computed from the original and locally normalized image. Note that C can be

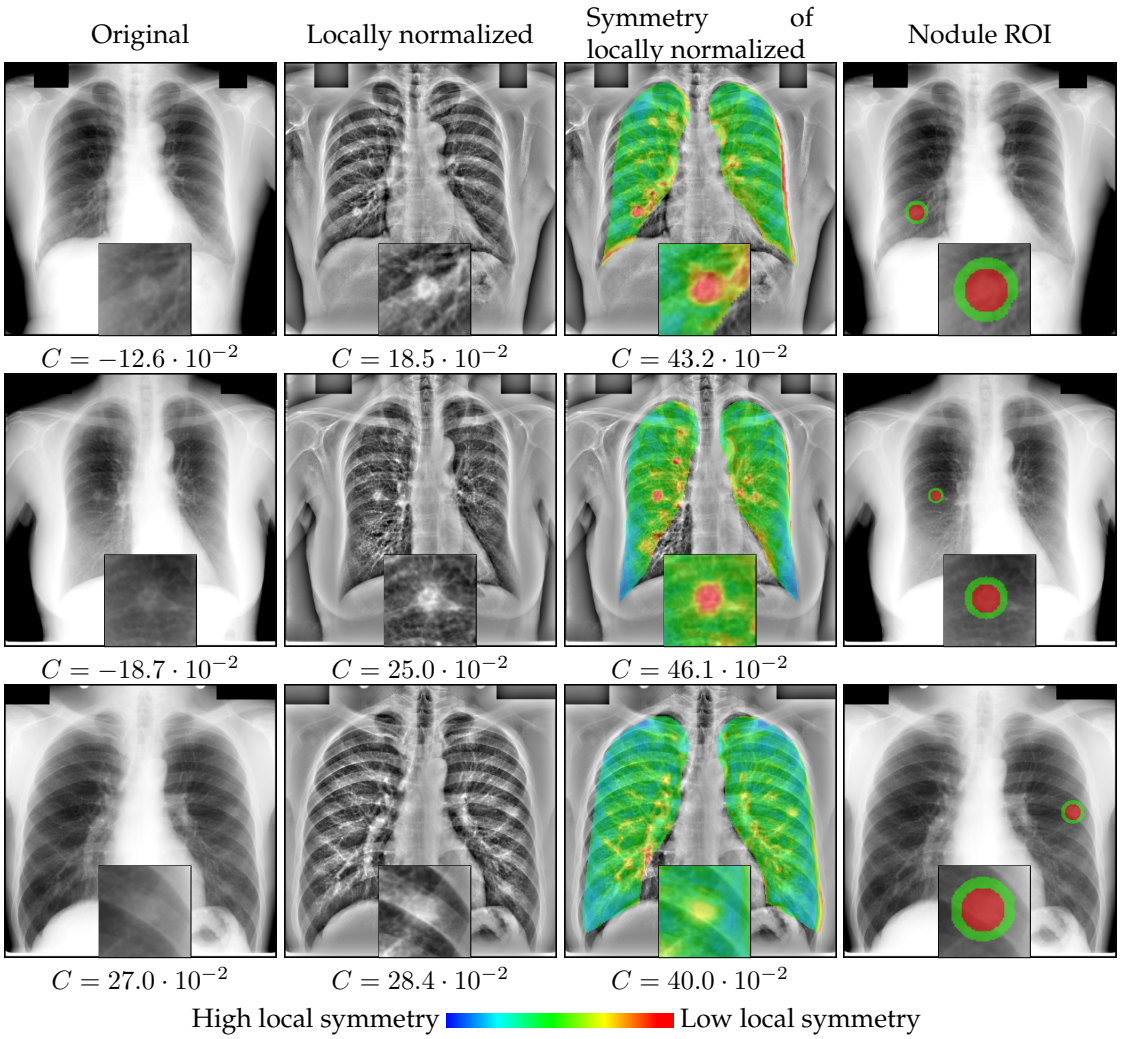
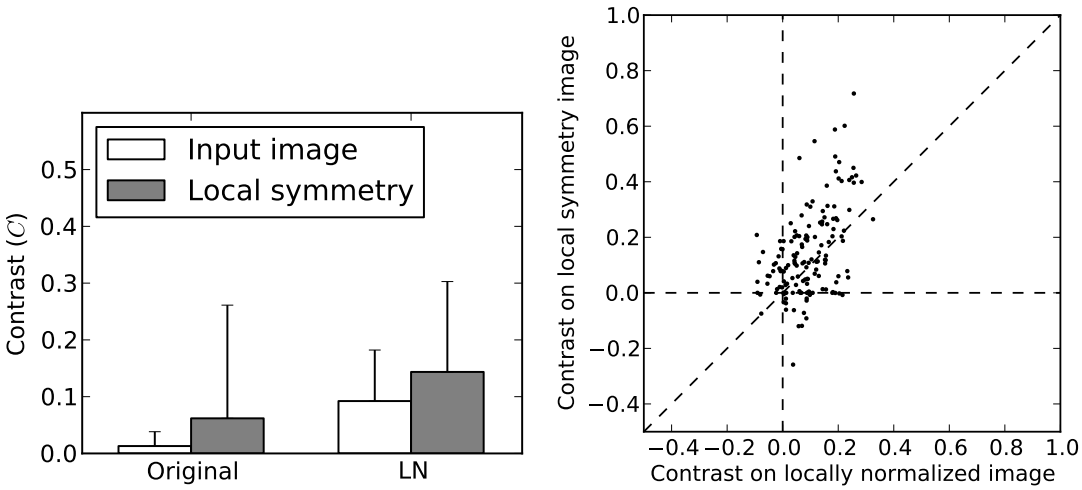


Figure 5.7: Local symmetry maps for three CXRs containing nodules. Shown are original images, locally normalized images, symmetry maps calculated from locally normalized images, and the nodule ROI (red) and background region (green) used for contrast computation. Insets show a detailed view of the nodule ROI. The nodule contrast C is indicated below the images.

negative when the surroundings have higher values than the nodule ROI.

Results

Fig. 5.7 shows three examples of original CXRs containing nodules, locally normalized images, local symmetry maps, and the nodule ROI. In the local symmetry maps, an increase of the values relative to the surroundings is observed at the nodule locations. Note the near absence of rib and clavicle patterns, which are one of the most prominent structures in CXRs, but do not show a pronounced response in the local symmetry map because they exhibit strong symmetry. At



(a) Contrast after local symmetry computation (b) Contrast change of individual nodules

Figure 5.8: Nodule contrast C measured in input images and local symmetry maps computed for 150 nodules. (a) Change of C measured in original and LN images. Errorbars indicate standard deviation. (b) Change in C per nodule on locally normalized images and on local symmetry images computed on LN images. Each marker indicates a nodule.

a number of locations in the image s is increased although no abnormalities are present. This can be observed near the pulmonary vessel tree, at some crossings of ribs and vessels and close to the lateral rib cage.

Fig. 5.8(a) shows the average nodule contrast C for original images and LN images and for local symmetry maps computed from original and LN images. The local normalization procedure on its own increased C, but local symmetry computation further enhanced the contrast. For both types of images C increased significantly in the local symmetry map: from $1.3 \cdot 10^{-2}$ to $6.8 \cdot 10^{-2}$ ($p < 0.001$; paired Student's t-test) when using original images and from $9.1 \cdot 10^{-2}$ to $14.1 \cdot 10^{-2}$ ($p < 0.001$; paired Student's t-test) when using LN images. Fig. 5.8(b) shows the changes in C per nodule computed on LN images and local symmetry maps computed from the LN image. For most cases an improvement of C is observed. In LN images the maximum contrast was $32.5 \cdot 10^{-2}$ and in local symmetry maps $71.8 \cdot 10^{-2}$.

5.4 Discussion

We have presented a method to compute local and global symmetry efficiently in 2D gray value images. In applications concerning detection of pathology in chest

radiographs global symmetry was found to be a strong indicator for the overall presence of abnormalities, and local symmetry was an informative measure for localizing abnormalities. The method provides three contributions to the field of automatic medical image analysis: (1) a novel continuous symmetry measure was developed, (2) to our knowledge this is the first method that uses global symmetry to detect images containing abnormalities, (3) the method addresses the issue of inherent and pathological asymmetry by combining gray value and position information to quantify symmetry.

One of the most outstanding results in this work is that global symmetry as a single unsupervised feature performed as well for detecting abnormal CXRs as a previously published supervised method, which analyzes CXR locally for abnormalities based on labeled training examples[112,148]. Several reasons explain this result. The most important one is that in symmetry computation the CXR is used as its own reference for determining what is normal and abnormal; an important observation also mentioned by Sun et al.[218]. In this way the problems of inherent (non-pathological) differences between CXRs from different individuals, due to anatomical (e.g. shape and appearance of the ribcage) and physiological (e.g. age), but also acquisition differences (e.g. scanner model), are addressed at the same time. This self-normalizing property holds for any type of medical image. A second reason is that global symmetry provides a robust holistic interpretation of the full radiograph without a need to determine what kind of patterns are present. In this sense the method shows similarities to the first phase of the radiologist's reading process, in which a very short impression of the *Gestalt* of the image already provides a first clue to the presence of abnormalities[227,228]. A disadvantage of summary statistics, such as global symmetry, is that they can only detect images with relatively large abnormalities.

We found that local symmetry was an informative feature, which improved detection of textural and small nodular abnormalities. When added to a set of texture features, it improved the detection of textural abnormalities compared to texture features alone. It might be surprising that adding local symmetry improved performance significantly in a combination with a large set (106) of other features. We hypothesize that a reason for this is the use of the nonlinear GentleBoost classifier. This type of classifier uses an implicit feature selection in each iteration of its training phase, where the feature is selected that minimizes the error for the current weighting of samples in the dataset using a (weak) regression stump classifier[58]. An illustration of the importance of local symmetry is that, in

the patch classification experiment, the feature was selected first – indicating that it has the highest discriminatory performance of all the features – and in total in 6 out of 100 regression stumps of GentleBoost. In the third experiment we found that the local symmetry enhances the contrast of nodules and in some cases even strongly highlighted the correct location of the nodule. This property could be used in automated nodule detection, in addition with other features, to improve detection and classification.

A property of symmetry computation is that, without adding additional information, it leads to symmetrical structure in the symmetry map. Especially when the computation is limited to spatially symmetrical positions, such as in a number of previously published methods[213,218,219], the symmetry map is by definition fully symmetric. In medical images, which often do not exhibit perfect symmetry, even in normal examples, this leads to an artificial elevation of symmetry measures. It will also lead to an ambiguity of the side where abnormalities, such as pathology, are located. We have addressed this issue by allowing flexibility in matching positions, using a weighting factor which controls the influence of position and local density patterns. Smaller abnormalities can then be unambiguously localized, such as is visible in Fig. 5.7 where nodules are highlighted on the correct side in the local symmetry map. In contrast, using only the exact reflected point for symmetry computation corresponds to very high position weights in the presented method. Our results showed that these higher weights have lower discriminative performance than lower weights.

In the field of automatic CXR analysis one of the few methods that uses left-/right symmetry directly is the contralateral subtraction technique by Li et al.[93,94]. After determination of the symmetry axis based on rib profiles, the axis is rotated upright using a minimization procedure. Then global and elastic registration were used to align the lung structures. Although ribs were visually determined to have been removed in the majority of images, no evaluation was performed of the method's value in pathology detection. An important difference with our method is that we do not attempt to solve the registration problem. Registration can be difficult, especially in pathological images, requires appropriate regularization, and leads to image artifacts[218]. In fact, there is no perfect registration possible between contralateral lung fields.

In generic 2D images a full search for the position and orientation of the symmetry axis is required as these parameters are unknown. Existing methods in literature have therefore included methods to find the axis as an integral part of

the algorithm[213,219]. In medical images it is often possible to make a good initial approximation based on the prior knowledge of the image content. Instead of requiring a full search, we refine this initial approximation, which can be usually accurately found because the danger of being trapped in a local minimum is small. Such an approach shares similarities to the work of Liu et al.[217], who used minimization techniques to determine the midsagittal plane in pathological brain MRI. We found that an optimal rotation of the image led to large performance increases of global symmetry compared to the original image. The optimization of the x-location of the symmetry axis led to only a minor performance increase. This is explained by the property that the algorithm is relatively insensitive to translations in the x-direction; minor position changes of corresponding patches will lead to overall slightly higher local symmetry values, but not to a loss of the discriminative properties of global symmetry.

In the case of chest radiographs the initial symmetry axis is determined from the lung segmentation. Although this requires the existence of a proper segmentation, the segmentation serves the additional purpose of excluding inherently asymmetrical parts of the image; the heart region in the case of CXRs. No information regarding symmetry is available in that region; for the discriminative properties of global symmetry this is an issue when abnormalities are only present in the excluded areas. In these excluded areas we set local symmetry values to a value of 0, in order to still allow local analysis by combining local symmetry with other local features. Alternatively, classification techniques dealing with missing values could be used.

There are a number of situations where the optimization of the symmetry axis can fail. The most prominent one is the presence of very large, unexpected, asymmetrical structures, such as gross pathology. The optimization can then be trapped in a local minimum, for example when the symmetry axis has been positioned in such a way that the pathology is aligned with a structure of similar (high) density on the other side. This failure to detect the correct upright position in abnormal images is not necessarily a problem, as global symmetry scores will remain high. A more difficult situation is the presence of density differences on the sides of the symmetry axis which are not caused by pathological processes. An example of such a situation is rotation of the rib cage around the caudocranial axis, which causes a slight intensity difference between left and right lung fields. This specific problem could be addressed by using contrast invariant descriptors for $\mathbf{f_I}$, but might lead to loss of performance in abnormal images where density

differences play an important role.

The optimal type of descriptor for use in symmetry computation could be further investigated, see Mikolajczyk et al.[229] for an overview of existing techniques. Of the point descriptors, SIFT descriptors are a popular point descriptor and they have been successfully employed in symmetry computation[213]. In this work we used raw patch values. This type of descriptor was shown to have similar performance as SIFT descriptors in a content based image retrieval application for the detection of abnormalities in CXRs[224]. One of the reasons why raw intensity values work well in chest radiographs is that rotation and scale invariance are not required. On the contrary, the orientation of certain anatomical structures, such as ribs, provide valuable information for symmetry computation. In other types of medical images it might be beneficial to explore different types of point descriptors and distance measures. Regarding distance measures one can think of cross-correlation to provide contrast invariance, and mutual information for nonlinear intensity relations between similar patches.

The presented method computes symmetry in 2D gray-scale images, but it can be easily extended to N-D images, such as CT and MRI, and color images, such as retinal or microscopic pathology images. The only requirement is that an appropriate point descriptor and similarity measure is used. Radial symmetric structures can be addressed in the same framework by a slight modification of the algorithm. Because the method is designed to deal with the presence of normal asymmetry, structures do not have to exhibit perfect symmetry to be suitable for analysis.

5.5 Conclusion

An efficient method to quantify local and global symmetry in medical images was presented. The method is designed to work under conditions of normal inherent asymmetry and pathology induced asymmetry. In three experiments on chest radiographs it was demonstrated that local and global symmetry are strong indicators for the presence of pathology.

Acknowledgments

The authors kindly thank Helen Ayles from the London School of Hygiene & Tropical Medicine and the Zambia AIDS-Related TB (ZAMBART) Project for providing the TB suspects dataset used in this study.

This study was supported by the European and Developing Countries Clinical Trials Partnership (EDCTP), the "Evaluation of multiple novel and emerging technologies for TB diagnosis, in smear-negative and HIV-infected persons, in high burden countries" (TB-NEAT) project.

Automatic detection of tuberculosis

6

Abstract

Tuberculosis (TB) is a common disease with high mortality and morbidity rates worldwide. The chest radiograph (CXR) is frequently used in diagnostic algorithms for pulmonary TB. Automatic systems to detect TB on CXRs can improve the efficiency of such diagnostic algorithms, however the diverse manifestation of TB on CXRs from different populations requires a system that can deal with different types of abnormalities.

We developed a computer aided detection (CAD) system which combines the results of supervised subsystems detecting textural, shape, and focal abnormalities into one TB score. The textural abnormality subsystem provided several subscores analyzing different types of textural abnormalities and different regions in the lung. The shape and focal abnormality subsystem each provided one subscore. In the combined system one overall TB score was computed by normalizing the subscores, collecting them in a feature vector, and then combining them using a supervised classifier.

Two databases, both consisting of 200 digital CXRs, were used for evaluation, acquired from (A) high-risk group screening in London, UK (Find & Treat) and (B) TB suspects in Capetown, South Africa (TB-NEAT). The subsystems and combined system were compared to two references: an external reference set by sputum culture and a radiological reference determined by a human expert. The area under the ROC curve A_z was used to measure performance. Additionally, the performance of an independent human observer was compared to the best individual subscore and to the combined system.

For database A, the focal lesion detector and the texture subscore measuring large opacities were the best performing subscores with A_z = 0.827 and 0.821 for the external and radiological reference respectively, whereas in database B the texture subscore measuring large opacities had the highest performance, with A_z = 0.759 and 0.866. The combined system performed better than the individual subscores, except for the external reference in database B, giving performances of 0.868 and 0.847 in database A and 0.741 and 0.899 in database B. The performances of the independent observer, 0.910 and 0.942 in the database A and 0.755 and 0.939 in database B, were slightly higher than the combined system. Compared to the external reference, differences in performance between the combined system and the independent observer were not significant in both databases.

The combined CAD system performed better than the individual subscores and approaches performance of human observers with respect to the external and radiological reference. The use of supervised combination to compute an overall TB score allows for easy adaptation to different types of settings and operational requirements.

6.1 Introduction

Tuberculosis (TB) remains one of the world's major health concerns. In 2011 an estimated 8.7 million new cases and 1.4 million deaths were reported. The majority of the TB burden is located in Africa, followed by the Asian countries[1]. Although the overall incidence of TB in the Western World has been decreasing in the past decades, an increase in TB rates has been reported in selected high-risk populations especially in urban settings[4,5]. Chest radiography is becoming increasingly important in the fight against TB, because existing screening diagnostics such as sputum staining have become less reliable in populations with a high prevalence of HIV/AIDS[10]. With the increasing availability of digital radiography[34], computer aided detection (CAD) systems can be developed that could facilitate mass population screening for TB. In high burden countries the number of skilled human CXR readers is often low, and the intended use of CAD in this study is as a first screening test that selects cases that require follow-up diagnostic tests such as sputum culture.

The pathofysiology of TB is complex[7,230] leading to a large diversity of pathologic changes in the lungs and other parts of the body. This diversity is reflected in a wide variety of pulmonary manifestations on the chest radiograph (CXR) with distinct morphological patterns[7]. Known causes for this diversity are age[38,39], ethnicity[40], and co-infection with HIV[41]. The presentation on CXR also differs with different stages of the disease. Traditionally, differences have been described between primary and post-primary TB[8], and in general TB in an early stage shows smaller, but also morphologically different, abnormalities than in more advanced stages of the disease. These variations lead to different frequencies of distinct patterns across populations. A generally applicable CAD system requires good performance for all the distinct patterns, but also a methodology to adapt it to individual populations with their own specific characteristics. In this study we focus on three categories of abnormalities: textural abnormalities, characterized by diffuse changes in appearance and structure of a region; focal abnormalities, which are isolated circumscribed changes in density; and shape abnormalities, where disease processes have altered the contour of normal anatomical structures.

Most of the previously developed systems for automatic analysis of chest radiographs have focused on single tasks: nodule detection[97], interstitial abnormalities[103,147,231,232], or lung shape abnormalities[105]. An extensive overview can

be found in Katsuragawa et al.[73] and van Ginneken et al.[74]. A recent overview of automatic TB detection in chest radiographs is given in Jaeger et al.[106]. Koeslag et al.[110] used template matching in the Fourier domain to determine the presence of miliary TB in an African setting. In a study by van Ginneken et al.[102] texture analysis was employed to classify chest radiographs, acquired in a TB screening program, as normal or abnormal. Arzhaeva et al.[233] reported on a different automatic system which classifies chest radiographs as normal or suspect for TB based on its global appearance. Hogeweg et al.[148] reported on an improved TB detection system, and evaluated it in a database of radiographs acquired from TB suspects in Africa, showing the benefit of a combination of local and global features of the radiograph. Tan et al.[115] analyzed distributions of intensities in interactively segmented lung fields in a dataset of TB cases from Southeast Asia. Jaeger et al.[116] computed a number of different texture feature sets in automatically segmented lungs to detect TB in a dataset obtained from a North American TB control program.

In this paper we propose an innovative combination of individual subsystems in a structured manner to address the problem of developing a CAD system with good generalization properties. For automatic TB detection there are two main reasons to combine systems. The first, mentioned before, is that it is not likely that a single system will suffice in a multitude of settings. A combination of multiple systems can be adapted to the specific setting, for example by weighing the output of a specific abnormality detection system higher when it is strongly associated with TB in a particular population. The second reason is a general beneficial effect of system combination on the performance of supervised systems. This effect has been extensively studied in the field of pattern recognition[234,235]. Niemeijer et al. showed that combination of independently developed systems, addressing the same task, improved the performance of CAD[62]. In contrast to that work, we propose a combination of heterogeneous systems, addressing different types of abnormalities.

In this paper textural, focal, and shape abnormality subsystems are combined into one system to deal with the heterogeneous abnormality expression in different populations. The performance is evaluated on a TB screening and a TB suspect database using both an external and a radiological reference standard. The paper is organized as follows. In Section 6.2 the systems to detect different types of TB related abnormalities and their combination is described and in Section 6.3 the two databases that were used for evaluation. Section 6.4 shows experiments

and results, which are discussed in Section 6.5, followed by the conclusions in Section 6.6.

6.2 Methods

The proposed combined CAD system consists of several subsystems, each of them producing one or more subscores indicating the presence of textural, focal, and shape abnormalities. All the subsystems are aggregated into one score by combination of the subscores. The subsystems depend on the segmentation of anatomical structures, preprocessing of the chest radiograph and computation of features, and these tasks are described first.

6.2.1 Segmentation

The lungs and clavicles were segmented to limit the analysis by the subsystems to the lung fields and provide them spatial context. The segmentation is based on supervised pixel classification and requires a set of features to be computed for each pixel.

Local feature computation

Local characteristics of each pixel in the image were computed. Three types of features were calculated: texture features based on Gaussian derivatives, features derived from the Hessian matrix, and position features. These features were computed at images resampled to a width of 256 pixels. To capture local image structure the output of Gaussian derivative filtered images of order 0 through 2 (L, L_x, L_y, L_{xx}, L_{xy}, L_{yy}), at scales 1, 2, 4, 8, 16 pixels were calculated[133]. Hessian matrix derived features, also calculated at scales 1, 2, 4, 8, and 16 pixels, were used to detect the presence of line like structures[135]. Considering the two eigenvalues of the Hessian matrix λ_1, λ_2, $|\lambda_1| > |\lambda_2|$ two measures were derived: (1) $\sqrt{(\lambda_1^2 - \lambda_2^2)}$ to extract the lineness of the local image structure, and (2) the largest absolute eigenvalue $|\lambda_1|$ to indicate the strength of the response. In addition two position features, the x- and y-coordinate normalized to the height of the image, were added. Each pixel is described by in total 43 local features.

Lung segmentation

A lung segmentation is required to limit the analysis to the region inside the lung fields, where TB primarily manifests itself. Two different segmentations were produced: (1) *lung-PC*, the post-processed output of a pixel classification stage

and (2) *lung-HAP*, which combines pixel classification and shape model information to improve segmentation of lungs containing gross abnormalities. *lung-PC* tends to classify grossly abnormal regions as non-lung and was later analyzed for shape abnormalities; whereas *lung-HAP* was used in the textural analysis and analyzes the full lung fields.

The pixel classification set-up for *lung-PC* is based on the method described in van Ginneken et al.[76]. The local features described in the previous section were computed for every pixel in the image. From a set of training images, examples of pixels in- and outside manually outlined lung fields were sampled (a random selection of 0.3% for both classes). A k-nearest-neighbor classifier ($k = 15$) was trained and used to assign to all pixels in a test image a lung likelihood p. The resulting lung likelihood map was converted to a binary segmentation in a series of steps: Gaussian blurring with $\sigma = 0.7$ pixels, thresholding at $p = 0.5$, selection of the two largest components, and morphologically closing with a spherical kernel of radius = 10 pixels.

The lung segmentation which includes shape information, *lung-HAP*, was provided by applying the **H**ybrid **A**ctive Shape Model **P**ixel Classification (HAP) algorithm[76,187]. The intensity model of the active shape model was trained on the likelihood map provided by pixel classification, instead of on the original image. The shape model was computed from the same training set as for the pixel classification stage.

Clavicle segmentation

The presence of many overlapping structures renders the upper lung region the most difficult to analyze in the chest radiograph and can lead to a high rate of false positives[148]. It is also the area where TB most commonly manifests itself[111]. The clavicle location was provided by the algorithm described in Hogeweg et al[187] (Chapter 3). In this method, supervised pixel classifiers were constructed to segment the interior, the head and the border of the clavicle. The local features described in Sect. 6.2.1 were used, the same as for the lung segmentation, together with context features extracted from the lung segmentation, namely the x- and y- coordinates normalized to the height of the bounding box of the lung fields, the distance to the lung wall and the distance to the center of gravity of the lung segmentation. Active shape model segmentation based on the interior segmentation was performed to generate an initial outline which was then refined using dynamic programming.

6.2.2 Preprocessing

Besides the lung fields, the CXR can contain other structures, such as parts of the abdomen, arms, and air outside the body. In a properly collimated image the proportional area of these structures is minimal. In practice there is considerable variation in what proportion of the total image is occupied by the lung fields. This variation was reduced by applying a virtual collimation procedure, which yields images with standardized lung sizes, improving robustness of subsequent analyses. The standardization of the scale intrinsic to the feature computations is important because normal structures in the CXR have typical sizes. A bounding box B around the lungs was determined from *lung-HAP* and the image was cropped to B. A 5% margin was added to the width and the height of B to compensate for possible undersegmentation and to reduce border effects in feature computation. The cropped image was then resized to a width of 1024 pixels.

6.2.3 Abnormality detection subsystems

A number of subsystems were used to detect textural, focal, and shape abnormalities in chest radiographs. One or more subscores were generated by each subsystem. The subscores were afterwards combined into one overall abnormality score. The different subsystems and subscores are summarized in Table 6.1. In this section a detailed description of the subsystems is provided. Training sets and testing procedures for each subsystem are described in Appendix 6.A.

Shape analysis

When large abnormalities close to the lung walls are present, the normal shape of the lungs is corrupted and difficult to determine, because of similar densities of abnormalities and extra-pulmonary structures. This causes the boundaries of the detected lung fields to be displaced with respect to the true lung boundary. Therefore, an abnormal shape of the projected lung fields indicates the presence of abnormalities and can be used to detect abnormal images.

A shape abnormality score was computed by comparing a shape representation extracted from *lung-PC* to a set of normal lung shapes. The set of normal lung shapes were extracted from the automatically computed *lung-PC* segmentation from a set of normal images. Rays were cast in equiangular directions from the centroid of the detected lung and the distance to the intersection with the boundary was recorded. This creates a feature vector of length n for each shape, with n the number of directions. For each lung 80 rays ($n = 80$) were cast to cre-

Subsystem	Subscore	Description
Shape analysis	*S-shape*	Measures deviations from normal lung field shape
Texture analysis	*S-texture*	Measures severity and extent of all textural abnormalities
Regional texture analysis		
	S-texture-central	Similar to *S-texture* but only measured in the central area
	S-texture-upper	*idem* for upper area
	S-texture-middle	*idem* middle area
	S-texture-lower	*idem* lower area
Regional texture asymmetry		
	S-texture-central-asymmetry	Difference in texture score between left and right central areas
	S-texture-upper-asymmetry	*idem* for upper areas
	S-texture-middle-asymmetry	*idem* for middle areas
	S-texture-lower-asymmetry	*idem* for lower areas
Texture analysis - small opacities	*S-texture-small*	Measures severity and extent of small opacities
Texture analysis - large opacities	*S-texture-large*	Measures severity and extent of large opacities
Texture analysis - consolidation	*S-texture-consolidation*	Measures severity and extent of consolidations
Focal lesion detection	*S-focal*	Measures total load of focal lesions

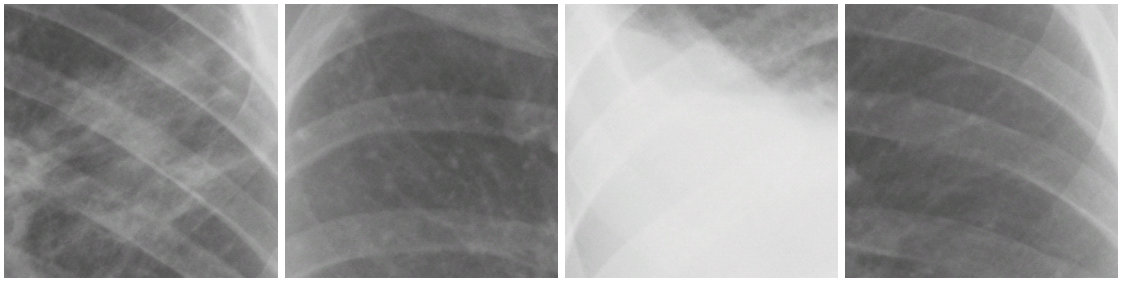
Table 6.1: Subsystems and subscores generated by subsystems

ate the shape. The feature vectors for the left and right lungs in the image were normalized to the height of the bounding box B and then concatenated to obtain one vector describing both lung shapes.

As the occurrence of lungs with an abnormal shape due to abnormalities adjacent to the lung wall is relatively uncommon and the abnormal shape is difficult to predict, a one-class classifier based on a PCA model of normal shapes (retaining 95% of the variance) was used to describe normal shapes and identify abnormal ones[236]. A large Mahalanobis distance (generalization of the standard score[54]) of a test shape to the model indicates a more abnormal shape. This Mahalanobis distance is used as *S-shape*.

Texture analysis

Textural abnormalities in CXRs commonly occur in TB and typically reflect inflammatory changes in the lung parenchyma, but can also be the result of fluid or fibrotic changes in the pleural space. The proposed texture analysis provides a likelihood of a textural abnormality being present for each pixel in the lung. The



(a) Large opacities (b) Small opacities (c) Consolidation (d) Normal

Figure 6.1: Examples of different types of abnormalities and normal appearance of the lung field. (a) Large opacities; a diffuse non uniform density increase (b) Small opacities; multiple small focal lesions with normal tissue in between (c) Consolidation; a uniform density increase (d) Normal; ribs and vascular structure are visible

detection of textural abnormalities is based on analysis of small circular image patches (radius $= 32$ pixels) sampled every 8 pixels.

Two sets of features were computed for each patch in the automatically segmented lung fields. Image characteristics were computed by extracting statistics of Gaussian derivative filtered of order 0 through 2 (L, L_x, L_y, L_{xx}, L_{xy}, L_{yy}), at scales 1, 2, 4, and 8 pixels. The first four moments (mean, standard deviation, skew, and kurtosis) of the intensity distribution of each Gaussian derivative filtered image and the original image were computed for each pixel inside its corresponding circular patch. In total, 100 patch features per pixel were computed. Additionally, spatial context features for each pixel were also calculated: the x- and y- coordinates (normalized to the height of the image), the x- and y-coordinates in the bounding box B, the distance to the lung wall, the distance to the center of gravity of the lung segmentation and the signed distance to the clavicle, which is positive outside the clavicle and negative inside. In total seven spatial context features were computed.

A GentleBoost classifier[58] was then trained with pixels from abnormal patches in abnormal lungs and normal pixels from normal lungs. GentleBoost used 100 regression stumps as weak classifiers. Only abnormal patches containing textural patterns, labeled as either large opacities, small opacities, or consolidation, were used. These categories of abnormalities are described in Section 6.3.2. In a test image, patches were sampled inside the segmented lung fields and classified. After classification, each patch in the test image was assigned a likelihood of being abnormal. The image score *S-texture* was computed from the cumulative distri-

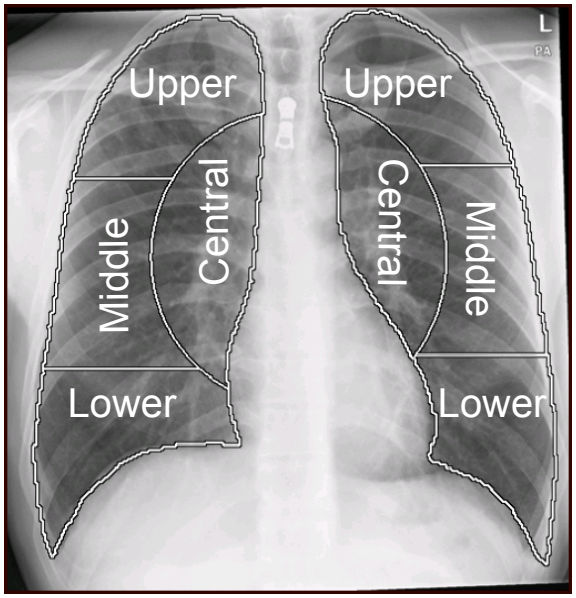


Figure 6.2: Division of automatically segmented lung fields into four subregions: central, upper, middle, and lower. Texture scores were computed from the four regions and from the difference in scores between left and right for each corresponding region.

bution of patch likelihoods by determining the likelihood corresponding to the 95% quantile. This score measures in a robust way the extent and the severity of the affected lung[149].

In addition to *S-texture*, which is one overall score for the whole lung fields and all types of textural abnormalities, separate scores were calculated for different regions of the lung and for different types of textural abnormalities. Regional texture scores were computed separately from subregions of the lungs. There are two reasons for adding regional texture scores: (1) a possible preferential location of TB in certain regions of the lung fields and (2) differences in performance of texture analysis in different regions. The subregions were based on a division in four parts of each lung field: upper, middle, lower, and central (Fig. 6.2). The central region was defined by a circle, with the centroid of the lung fields as its center, covering one quarter of the area of one lung field; whereas the upper, middle, and lower areas were defined by a vertical division in three equal parts of the remaining area. The corresponding scores *S-texture-central*, *S-texture-upper*, *S-texture-middle*, and *S-texture-lower* were calculated by converting the patch likelihoods inside each region to one score using the 95% quantile rule. In grossly abnormal cases, abnormalities can be present with roughly equal severity in both lung fields, but often, in more subtle cases, abnormalities are limited to one lung

field. This information was captured by computing the difference in texture scores between corresponding subregions on the left and right, giving *S-texture-central-asymmetry*, *S-texture-upper-asymmetry*, *S-texture-middle-asymmetry*, and *S-texture-lower-asymmetry*.

TB can cause multiple types of textural abnormalities with different visual patterns in chest radiographs namely large opacities, small opacities and consolidations. Examples of the patterns of these abnormalities are shown in Fig. 6.1. The same training procedure was followed as for general texture analysis but as positive training examples only patches from abnormal areas labeled with the specific type of abnormalities were used. The manual annotation of each abnormality is described in Section 6.3.2. After classification and the use of the quantile rule as previously described, three scores were computed: *S-texture-large*, *S-texture-small*, and *S-texture-consolidation*.

Focal lesion detection

Isolated well defined focal lesions, such as nodules, can occur in TB cases. This type of lesion is less well detected by texture analysis. Focal lesions were automatically detected with a commercially available software package for nodule detection (ClearRead+Detect v5.2; Riverain Technologies, Miamisburg, Ohio). The software outputs for each image a list of suspicious locations with a likelihood score. The total load of focal lesions was summarized into one image score *S-focal* by summation of the likelihood scores of all detected lesions.

6.2.4 Combination

With all the subsystems and subscores available the key issue is how to combine this information into one score that reflects the overall probability of the image containing abnormalities related to TB. From each individual subsystem the subscores S_i, with $i = 1, ..., N$ and N the total number of subscores, are collected into a vector $\mathbf{s}$. The 14 subscores were combined into one final score S_c for each case by classifying $\mathbf{s}$. Combination of $\mathbf{s}$ can be performed using either a static rule[235], or a learned (supervised) rule[237]. The basic difference between these two approaches is that static combination does not use image labels, while supervised classification does.

The subscores produced by the subsystems lie in different ranges and have to be normalized before they are combined, in order to weigh subsystems equally before the combination rule is applied. Subscores were normalized by transform-

ing them to zero mean and unit standard deviation. The normalization parameters were computed from subscores in a training set.

In static combination S_c is computed from $\mathbf{s}$ directly, without requiring a reference training set. In this paper, we investigated three different static rules: the sum rule $S_c = \sum_{i=1}^{N} S_i$, the product rule $S_c = \prod_{i=1}^{N} S_i$, and the maximum rule $S_c = \max_i S_i$. In supervised combination S_c is determined by classifying $\mathbf{s}$ using a learned classifier. For supervised combination linear discriminant analysis (LDA)[54], k-nearest-neighbor (kNN)[54], GentleBoost[58], and a Random Forest[59] classifier were used. Details about the classifier settings are given in Appendix 6.B. The trained classifier is constructed from a labeled training dataset. The image label is provided by a reference for every $\mathbf{s}$ in the training set; in this work we report results using two different reference standards, one based on expert reading of the radiograph, and one an external reference. The reference standards are described in Section 6.3.2.

6.3 Materials

Two datasets one from a European and one from an African country with different populations and settings were used to evaluate the CAD system. For each dataset an external and radiological reference standard were available.

6.3.1 Evaluation datasets

The first database is a set of 200 digital CXR (DigitalDiagnost Trixel; Philips Healthcare, The Netherlands) from the Find & Treat screening program. The Find & Treat program is aimed at screening for tuberculosis in high-risk groups in London, United Kingdom[20]. The high-risk population consists mainly of homeless people, prisoners, and problem drug users. In a period of 5 years (2005-2010), 104 active TB cases were found by the program. Active TB cases were defined as cases where a clinical decision was made to start TB treatment, in most cases based on a positive sputum culture test. The CXR have a isotropic pixel spacing of 143 μm and image widths in the range of 1800-3000 pixels. We selected 87 active TB and added 113 randomly chosen normal cases.

The second database consists of a set of 200 digital CXR (EasyDR; Delft Imaging Systems, The Netherlands) acquired from the Cape Town site of the TB-NEAT research study[238]. This study evaluates multiple diagnostics for TB in high burden countries. The radiographs have a pixel spacing of 250 μm isotropic and image widths in the range of 1500-1800 pixels. For all cases sputum culture results

were available. We selected 66 culture positive cases and added 134 randomly chosen culture negative cases.

6.3.2 Reference standard

For each database external and radiological reference standards were provided and used for two purposes: (1) to evaluate the performance of individual subscores, as well as the combined system and the performance of a second independent human observer; and (2) to train supervised systems. The external reference standard for tuberculosis was set by an independent test not associated with the CXR; the result of sputum culture testing for the TB-NEAT database and a combination of sputum culture testing and clinical diagnosis for the Find & Treat database. Sputum culture is considered the most accurate diagnostic test for TB and is typically used as reference standard in evaluation studies of other diagnostics[10,13]. The radiological reference standard was set by an experienced chest radiologist. Images were scored based on (1) the presence of any abnormalities and (2) the presence of abnormalities consistent with tuberculosis. These scores, ranging from 0 to 100, express the observer's certainty about the presence of abnormalities. A score of 50 or higher for any abnormalities corresponds to a radiologically abnormal case. A second chest radiograph recording system (CRRS)[42] certified reader independently scored the CXRs as well.

The training set for texture subsystems was created by outlining abnormal regions in all images by a third observer. Images were annotated using an extended version of the CRRS system. The CRRS categories large opacities and small opacities were annotated. A description of these categories can be found in[42,239]. Furthermore, we added an extra consolidation category, defined as an area of homogeneous density increase and considered as separate from the existing CRRS categories.

6.4 Experiments & results

Training, testing and analyses of all components was performed on the Find & Treat and TB-NEAT dataset individually. Training sets and test procedures are described in detail in Appendix 6.A. Receiver Operating Characteristic (ROC) analysis was performed and performance measures are given as the Area under the ROC curve (A_z), unless otherwise indicated.

	Database	Find & Treat		TB-NEAT	
	Evaluation reference	External	Radiological	External	Radiological
Subscores	S-shape	0.611	0.610	0.554	0.693
	S-texture	0.783	0.777	<u>0.759</u>	<u>0.866</u>
	S-texture-central	0.750	0.759	0.738	0.855
	S-texture-upper	0.784	0.787	0.696	0.796
	S-texture-middle	0.717	0.697	0.754	0.854
	S-texture-lower	0.474	0.455	0.570	0.645
	S-texture-central-asymmetry	0.759	0.733	0.705	0.781
	S-texture-upper-asymmetry	0.698	0.680	0.625	0.673
	S-texture-middle-asymmetry	0.710	0.671	0.696	0.799
	S-texture-lower-asymmetry	0.495	0.470	0.590	0.691
	S-texture-small	0.712	0.702	0.724	0.830
	S-texture-large	0.823	<u>0.821</u>	0.748	0.861
	S-texture-consolidation	0.681	0.665	0.673	0.767
	S-focal	<u>0.827</u>	0.814	0.685	0.858
	Combined system	**0.868**	**0.847**	0.741	**0.899**
	Independent observer	0.910	0.942	0.755	0.939

Table 6.2: Performance (Area under the ROC curve) of individual subscores, combined system, and the independent observer. Combinations performing better than the best individual subscore are indicated in **bold**. The best individual subscore for the external and radiological reference is <u>underlined</u>. See Fig. 6.3 for a visual display of the values.

6.4.1 Performance of subsystems

The upper part of table 6.2 shows the performance of the individual subscores for the Find & Treat and TB-NEAT database. The performance is shown for the external and radiological evaluation reference. For the Find & Treat database *S-focal* and *S-texture-large* are the best performing subscores, achieving A_z values of 0.827 and 0.821 respectively; whereas for the TB-NEAT database *S-texture* was the best subscore for both references, achieving A_z values of 0.759 and 0.866, respectively. An important observation is that subsystems perform differently in both databases, as visually illustrated in Fig. 6.3. For example, *S-texture* had highest performance in the TB-NEAT database for the external reference, but lower in the Find & Treat database, where *S-focal* performed better. Also the performance in the upper lung fields is higher than in the lower lung fields for both databases and references, a finding consistent with the known preference of TB for the upper lung fields.

Fig. 6.4 shows the likelihood maps of the texture analysis and focal lesion detection system for two examples of abnormal cases. The Find & Treat case is an example of a case where texture analysis gave a low score, but focal lesion detection a high score. The TB-NEAT case is an example where texture analysis

gave a high score, but no focal lesions were found. These examples illustrate the importance of detecting multiple types of abnormalities.

6.4.2 Combination

The combination of subsystems was performed using four supervised and three unsupervised rules. A summary of the results obtained with the different combination strategies are presented in Appendix 6.B. Supervised classification with the Random Forest classifier had on average the highest performance across databases and the two evaluation references. These combination settings were therefore used in all subsequent experiments.

Table 6.2 and Fig. 6.3 compare the performances of the subsystems, combined system, and the independent observer between the Find & Treat and TB-NEAT database. The combined system outperformed the subsystems for both references and databases. In the Find & Treat database the combined system achieved $A_z = 0.868$ and 0.847 for the external and radiological reference, respectively. In the TB-NEAT database the combined system achieved $A_z = 0.741$ and 0.899 for the external and radiological reference, respectively.

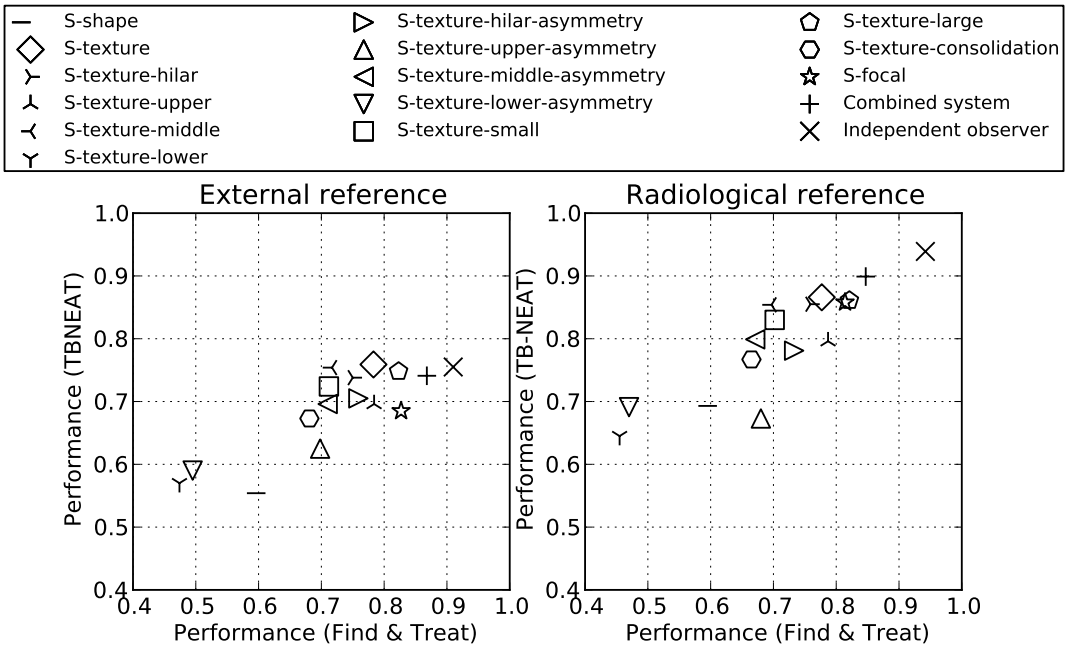


Figure 6.3: Scatterplots comparing performance of subscores, combination, and independent observer between the Find & Treat and TB-NEAT database. Left: external reference, right: radiological reference. See Table 6.2 for the corresponding values.

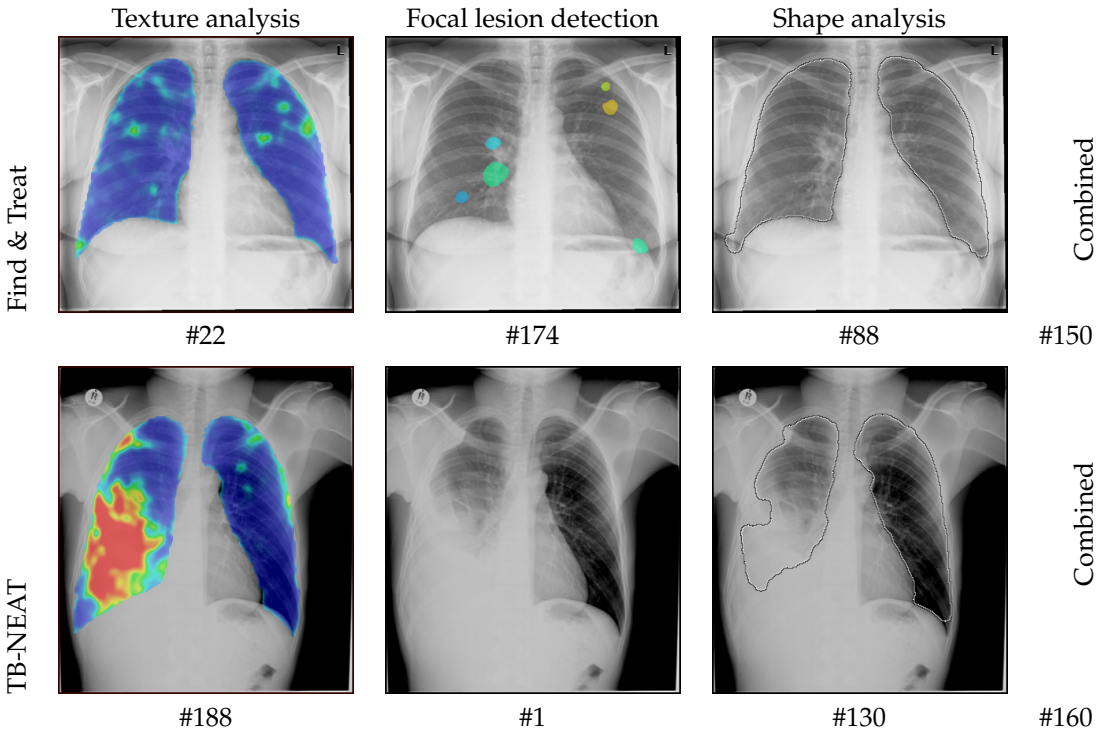


Figure 6.4: Output of texture analysis, focal lesion detection, and shape analysis for two cases with different types of abnormalities. Both cases were abnormal according to both the external and radiological reference standards. Colors indicate suspiciousness in the order blue-green-yellow-orange-red. The outline of the pixel classification based lung segmentation is shown for shape analysis. The numbers below the images indicate the position among the ranked scores of the system, where #1 is the most normal and #200 the most abnormal image. The last column indicates the ranked position in the combined system.

Fig. 6.5 shows for both databases and the external and radiological reference standard the ROC curves of the combined system, the best individual subscore per database/reference pair, and the independent observer. Using the external reference standard, we can compute the sensitivity and specificity of the reference observer for detection of TB, which is indicated as a single cut-off point in the ROC curve. In the Find & Treat database the reference observer made only a few false positive decisions, indicated by the high specificity of 98%, but with a relatively low sensitivity of 82%. In the TB-NEAT database the reference observer operated at a specificity of 54% and a sensitivity of 89%.

Fig. 6.6 and 6.7 visually show results of image classification using the combined system for the Find & Treat and TB-NEAT database, respectively. The first two rows show respectively the most abnormal and most normal images according to the combined system, i.e. the overall highest and lowest scoring images.

The third row shows the most prominent false positives, the radiological negative images with the highest scores with respect to the radiological reference. The fourth row shows the most prominent false negatives, the radiological positive images with the lowest scores. The degree of abnormality of an image is expressed by its ranked position among the scores of the 200 evaluation images (#1 most normal, #200 most abnormal) and also by the (*false positive fraction,true positive fraction*) position on the ROC curve.

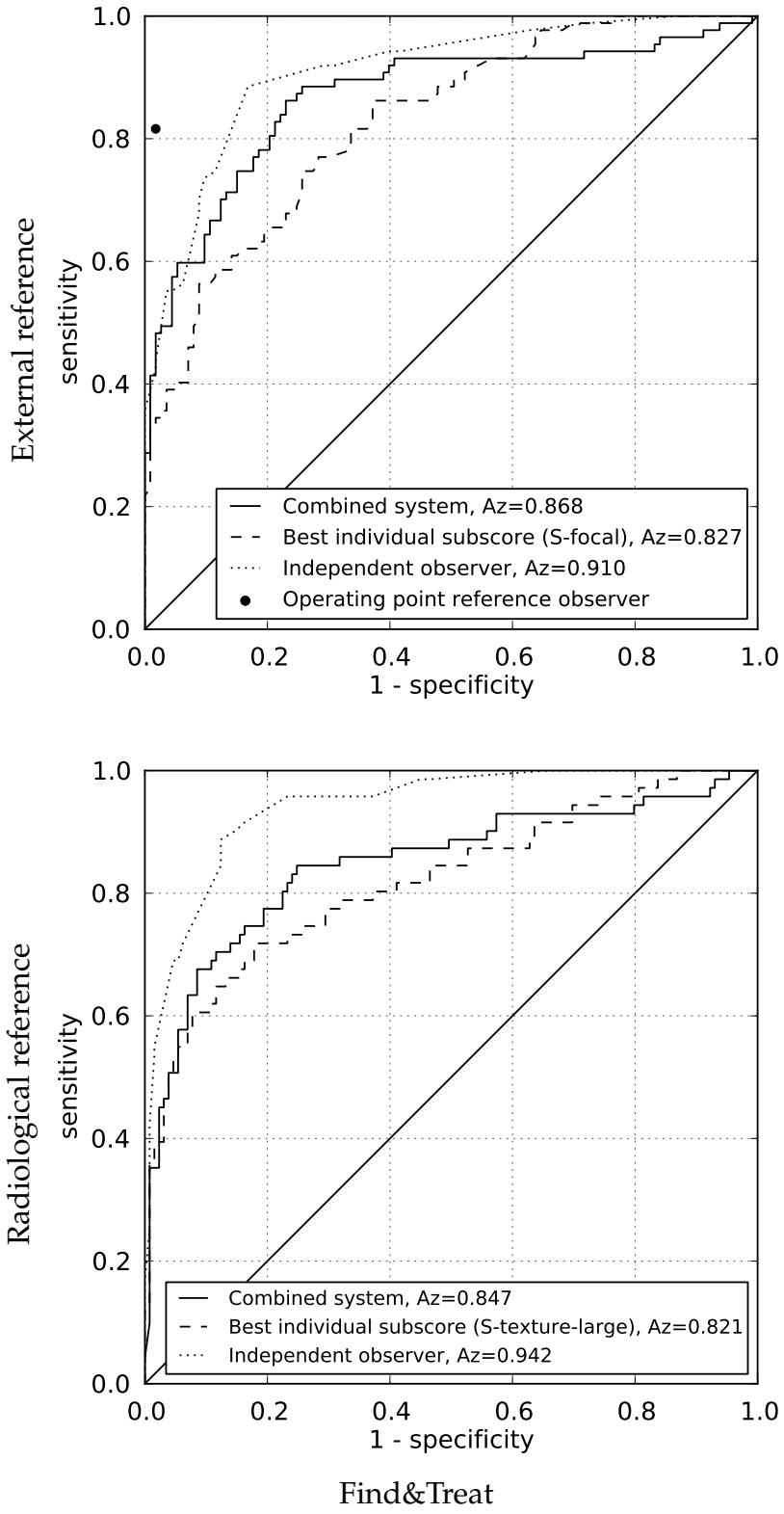


Figure 6.5: ROC curves for combined system, best subscore, and the independent observer. The upper ROC shows the results evaluated using the external (culture) reference, the bottom ROC using the radiological reference. This page: Find & Treat, next page TB-NEAT. For the external reference also the operating point, the threshold for normal/abnormal, of the reference observer is indicated.

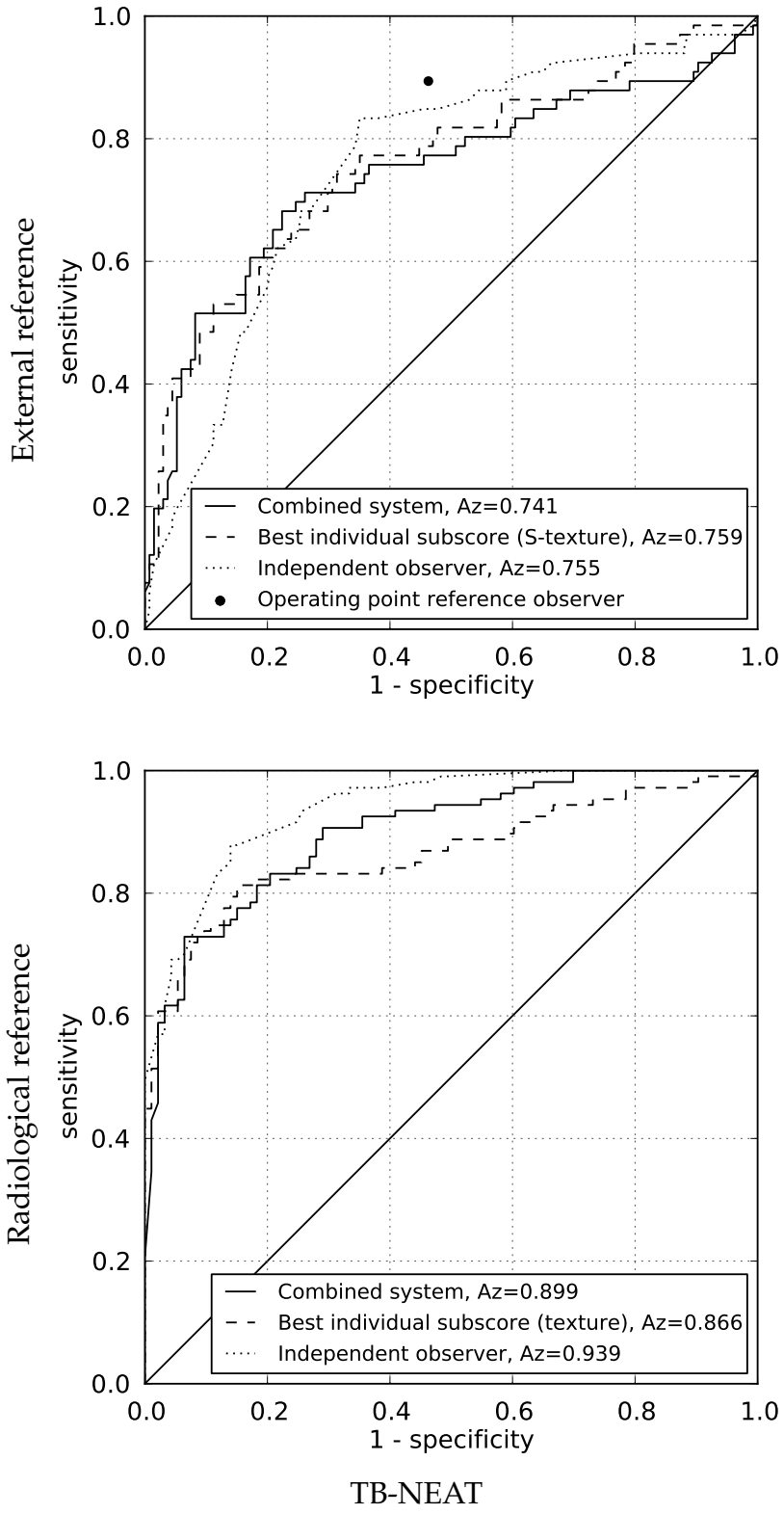


Figure 6.5: *continued*

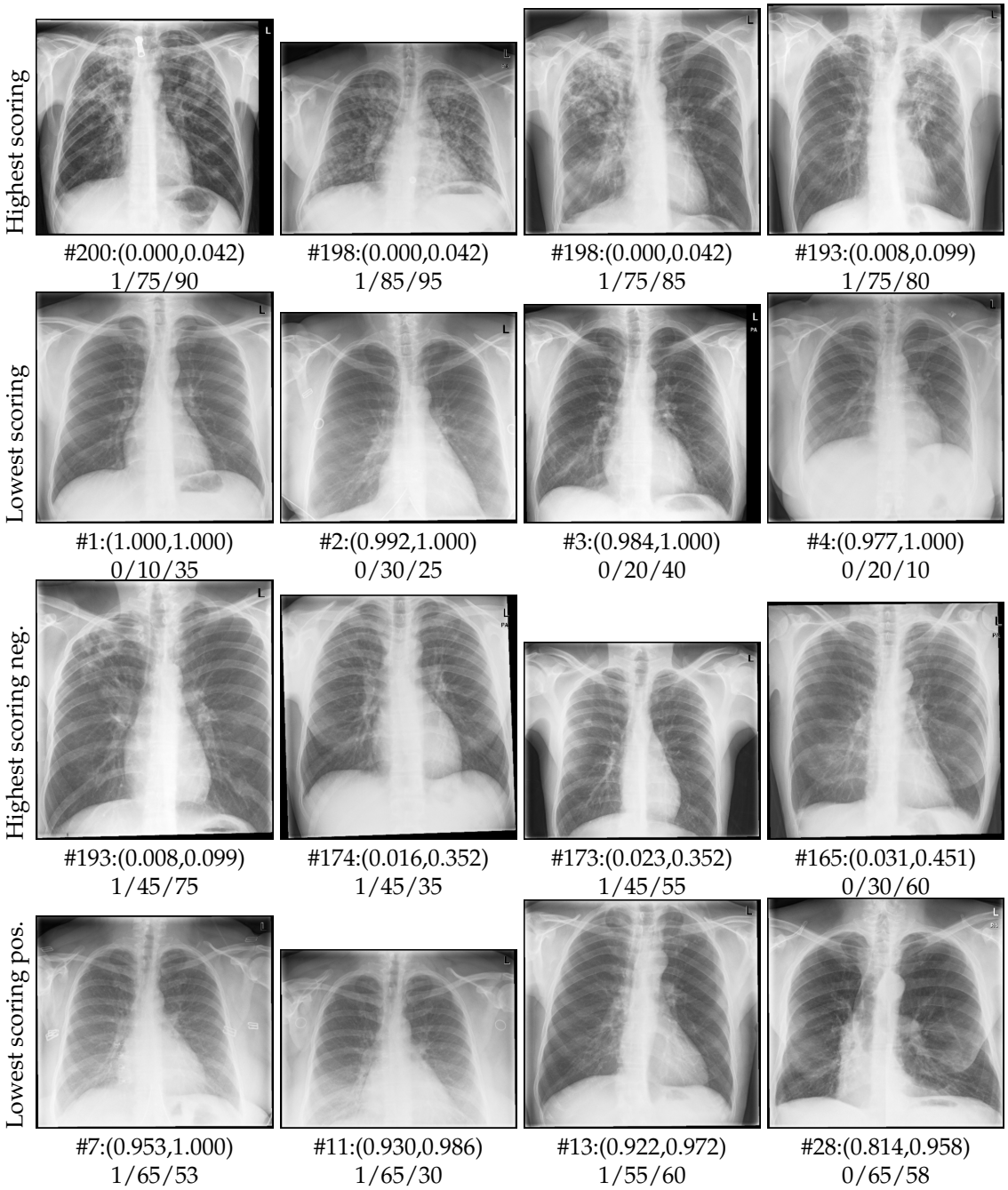


Figure 6.6: Visual summary of classification results on the Find & Treat database using the combined system. The first two rows show the overall highest and lowest scoring images. The third row and fourth row show the most difficult cases, respectively the highest scoring negative images and lowest scoring positive images according to the radiological reference. Numbers below the image indicate "rank:(FPF,TPF)/external reference/score reference observer/score independent observer", rank = sorted position among 200 evaluation cases (low = normal, high = abnormal), (FPF,TPF) indicates position on the ROC curve, FPF = false positive fraction, TPF = true positive fraction.

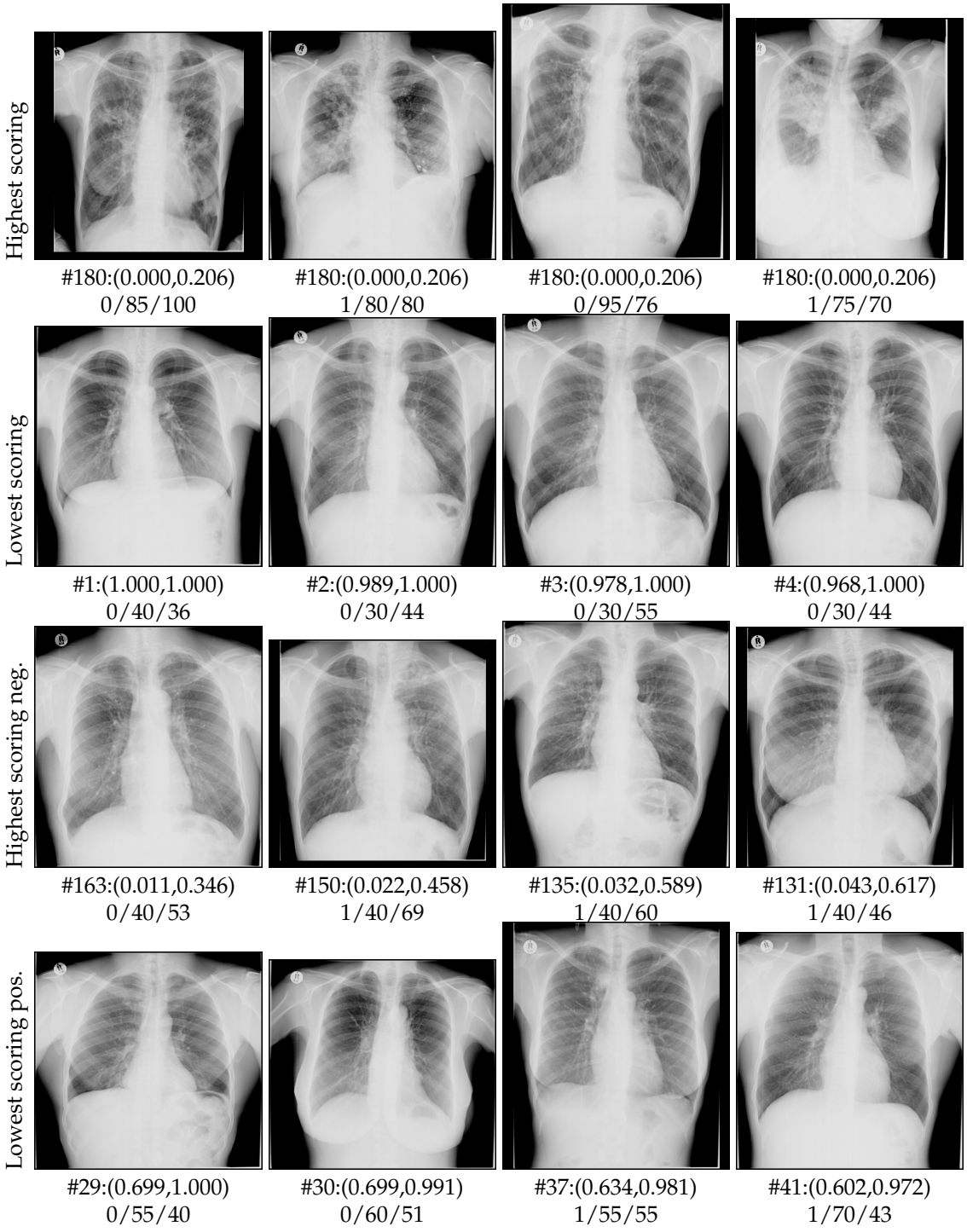


Figure 6.7: Visual summary of classification results on the TB-NEAT database using the combined system. See Fig. 6.6 for an explanation.

6.4.3 Automatic analysis versus reading by humans

The performance of the independent human observer was compared to the individual subscores and the combined system for both reference standards in order to judge the value of automatic analysis as a reader. For comparison with the external and radiological reference, the observer's score for abnormalities consistent with TB and the score for any abnormalities were used, respectively. For both databases and references, except for the external reference of TB-NEAT, the independent observer performed better than the individual subscores and slightly better than the combined system. Statistical comparisons of performances were made using case-based bootstrapping[68]. Differences were considered significant at $\alpha = 0.05$. For the Find & Treat dataset and compared to the radiological reference the independent observer performed significantly better than *S-texture-large* ($p < 0.001$), and better than the combined CAD system ($p < 0.001$). Compared to the external reference the independent observer performed again better than *S-focal* ($p = 0.001$), and also better than the combined system, but not significantly ($p = 0.055$). In the TB-NEAT database and compared to the radiological reference, there is again a significant difference between the independent observer and the best subscore *S-texture* ($p < 0.001$) and between the independent observer and the combined system ($p = 0.008$). For the external reference the independent observer had similar performance as *S-texture* and the combined system ($p = 0.446$ and $p = 0.351$ respectively).

6.5 Discussion

TB has a diverse presentation on chest radiographs and a multitude of differently appearing abnormalities have to be detected in order to achieve consistent performance in different populations. Three subsystems for TB detection and a framework to combine them into one overall score were presented. The contribution of this study in the field of automated pathology detection is two-fold: (1) differences in the performances of subsystems between datasets are accounted for, and (2) overall performance is increased. Specifically for TB detection, we have presented a technique to correct for differences in image collimation and provided methods to detect shape and textural abnormalities. The combined system obtained a consistent performance in populations with different TB representation, in contrast to the variable performance obtained by individual subsystems. We have shown that combination leads to improved performance compared to the

individual subsystems and approaches performance levels of human observers on two independent datasets.

Instead of using multiple detection subsystems to detect TB in chest radiographs it might be possible to design one comprehensive detection system. However, there are a number of reasons to prefer multiple small systems over one large complex system. The ability of a classification system to generalize is reduced when its complexity increases[240], requiring larger amounts of training data to achieve good performance. From a perspective of system design, multiple specialized subsystems are also preferable since these can be more easily tested and evaluated than larger general ones. A practical reason to use multiple subsystems, such as the focal and textural abnormality detection system, is that they can be developed parallelly by different research groups.

In CAD research it is common for studies to report that different types of systems do not perform consistently in the same task because the evaluation dataset was acquired in a different setting. We also observed this effect: with respect to the external reference the focal lesion detector had the highest performance in the Find & Treat database, but it is one of the lowest performing systems for the TB-NEAT database. Conversely, texture analysis had the highest performance in the TB-NEAT database, but a lower performance in Find & Treat. The performance of the combined system was higher or similar than the best individual subscores for both databases. This shows that, given an appropriate combination rule, selection of one subsystem out of a set of subsystems is not required. This property provides easy adaptation to a new setting where the relative incidence of abnormalities is different; instead of redesigning the CAD system, retraining the combination rule may be sufficient. Additionally, retraining the rule could be used to focus the CAD system on different tasks, e.g. detecting all abnormalities or specific types related to TB.

An interesting possibility would be to use the CAD system to discriminate between subgroups of abnormal CXRs. The first step is to discriminate active TB from other lung diseases such as other Mycobacterium strains[241], pneumoconiosis (coal miner's disease) and pneumonia. The next step is to discriminate inactive (old) TB from active TB. There is also a known variation in radiological appearance of active tuberculosis with certain patient characteristics, such as co-infection with HIV[41], ethnicity[40], and age[38]. We hypothesize that the variations in radiological manifestation due to different diseases and different patient characteristics should express themselves as differences between descriptors of the

radiological appearance. The feature vectors that were constructed in this paper provide such descriptors. Manual scoring systems for tuberculosis on CXR also provide feature vectors, but suffer from disagreement between readers, while the objective quantitative measurements provided by a CAD system do not have this limitation.

In the TB-NEAT database we found a considerable disagreement between the external and the radiological reference standard, which was also expressed by the large differences in performance of the CAD system. One reason for this difference is the uncertainty in the radiological reading. This is illustrated by the results in Table 6.3, where confusion matrices show that human observers make different kinds of errors with respect to the external reference. Therefore, in cases with observer scores closer to the binary cut-off value of 50 it becomes more uncertain whether the radiograph is truly abnormal. This effect is illustrated by the higher observer scores for culture positive cases (76 ± 11) than culture negative cases (71 ± 11) in the group of radiologically positive cases. Another reason for the discrepancy could be that it is known that the sensitivity of sputum culture is not 100%[242] and therefore some of the radiologically positive culture negative cases can be true TB cases. Another possibility is that these cases have abnormal X-rays due to diseases other than TB. The discordance between radiological and TB status also may explain why the combined system showed no improvement with respect to the best subscore.

Examining the most prominent mistakes of the combined system can provide insight into where improvement could be achieved. Such an analysis was made in Fig. 6.6 and 6.7 for the radiological reference. In the Find & Treat database three out of four highest scoring negative, i.e. false positive (FP), images, were positive for active TB, indicating that the combined system correctly judged them as suspicious. The first (ranked #193) looks clearly abnormal; a label of being radiological abnormal, which was given by the independent observer, would have been appropriate here. The other three cases are not overtly abnormal, but their lung fields display a more busy aspect than the overall lowest scoring cases. Of the lowest scoring positive, i.e. false negative (FN), three out of four images (ranked #7, #11, and #13) contain a pattern of small nodules, which was not well detected by the combined system. The case ranked #4 displays lymphadenopathy in the mediastinum and left hilus, which is a relatively uncommon and often subtle type of abnormality in TB. Adding subsystems specifically developed for small nodules and lymphadenopathy is expected to improve performance for images

(a) Find & Treat: reference observer

EXT \ RAD	Positive	Negative	Total
Positive	71	16	87
Negative	2	111	113
Total	71	119	200

(b) TB-NEAT: reference observer

EXT \ RAD	Positive	Negative	Total
Positive	59	7	66
Negative	62	72	134
Total	111	89	200

(c) Find & Treat: independent observer

EXT \ RAD	Positive	Negative	Total
Positive	64	23	87
Negative	11	102	113
Total	75	125	200

(d) TB-NEAT: independent observer

EXT \ RAD	Positive	Negative	Total
Positive	58	8	66
Negative	75	59	134
Total	133	67	200

Table 6.3: Agreement between observers (*RAD*) and external reference standard (*EXT*). Left: Find & Treat, right: TB-NEAT. Top: reference observer, bottom: independent observer.

containing these abnormalities. In the TB-NEAT database three out of four FP cases were active TB cases and these images are visually different than most normal images, which lead the independent observer to read two of these cases as radiologically abnormal. The TB case with rank #131 shows no clear abnormalities but has an unusually dense breast shadow, which has lead to false positive responses in texture analysis. Two of the FN images had no clear abnormalities, the case ranked #30 has a subtle abnormality in the upper left lobe and the case ranked #37 a subtle pattern of small nodules. The detection of such subtle abnormalities might benefit from an increase in the number of training examples.

For the external reference we found no significant differences between the combined system and the independent observer. This finding enables a potential replacement of human readers in certain screening algorithms, for example if the CXR is used to select cases that subsequently receive a confirmatory diagnostic test for TB. We are aware that, to conclusively show non-inferiority of CAD to human reading, larger datasets are required. This might also lead to an increase of the performance of the CAD system. Compared to the radiological reference the independent observer still performed better than the combined system. The main reason is a lower performance for detecting small and uncommon types of abnormalities. Nevertheless, CAD does not have to perform equally well as a human expert in all situations to be useful in practice. Depending on operational requirements and constraints – e.g. clinical practice, screening, or available funds – a slightly lower performance might be sufficient to for example reduce reading costs, or to provide quality assurance as a second reader.

Results showed that combined systems had higher performances than individual subscores. We explain the performance increase by the generally observed beneficial effect of combination on the performance of classification systems[234,235]. Further improvement in automatic detection of TB can be expected by adding results of other subsystems into the combined system. These systems can be different from the ones in our paper in terms of type of features, training sets, or preferably they follow a different algorithmic approach[103,224]. Their purpose can also be to detect other types of abnormalities, such as pleural effusions[224,243]. An important reason for combining systems which use a different approach is that the expected performance gain is higher when system subscores and errors are less correlated[60]. This explains the large benefit of adding the focal lesion detector which was found in the Find & Treat database. Another possibility to add more information to the combined system is by deriving multiple subscores from subsystems for example by calculating regional subscores, as was done in this study, or by including variations of the same subsystem.

6.6 Conclusion

TB has diverse manifestations and to analyze CXRs automatically, algorithms that focus on different manifestations need to be combined. When using only a single subsystem, different subsystems were found to perform best for two datasets from different populations, but the combined approach outperforms each single system in both cases. The combined system is close in performance to an independent human observer. Although the system presented combines multiple detectors, certain types of abnormalities are still missed. These can be addressed using the general framework proposed in this paper, where adding more subsystems is expected to further improve the versatility of an automated detection system.

Acknowledgments

This study was supported by the European and Developing Countries Clinical Trials Partnership (EDCTP), the "Evaluation of multiple novel and emerging technologies for TB diagnosis, in smear-negative and HIV-infected persons, in high burden countries" (TB-NEAT) project.

We would like to acknowledge the work of Jane Knight and Diana Taubman, the two reporting radiographers on the mobile X-ray unit in London who col-

lected all of the CXRs. We thank Alistair Story and Robert Aldridge for their help in collecting the Find & Treat database.

We thank Keertan Dheda, Rodney Dawson, and Grant Theron for providing us access to the TB-NEAT database and for their support in writing this chapter.

Finally we would like to thank Riverain Technologies for making a research version of ClearRead+Detect available for this study.

Appendix

6.A Training and testing of CAD components

Most subsystems of the CAD system require training data to perform their task. Table 6.4 shows the training and testing datasets for the components and Table 6.5 the number of images in the training sets. The two evaluation sets (*D-eval*) consisted of 200 images for both the Find & Treat and TB-NEAT database. Training sets for anatomy segmentation (*D-lung* and *D-clavicle*) were independent of *D-eval*. In the experiments with the Find & Treat database 495 consecutive images from the full screening database were used for lung segmentation. Most of these images were normal because of the low prevalence of abnormalities in a screening setting. A subset of *D-lung* consisting of 250 images was used for clavicle segmentation. In the experiments with the TB-NEAT database we used the same database as in[187], consisting of 548 images which had both lungs and clavicles annotated. This dataset is publicly available on `http://crass12.grand-challenge.org/`. In this set 225 and 333 images were considered normal and abnormal, respectively.

Shape analysis was trained with *D-shape*, a selection of normal images from *D-lung*. Because of the limited amount of data in *D-eval*, instead of splitting it into a training and a test dataset, 10-fold crossvalidation was used for texture analysis, score normalization and supervised combination. In each fold 9/10th of the cases were used for training or computation of normalization parameters, which was used for classification or normalization of the remaining 1/10th of test cases. Note that the reference standard used for training the supervised combination is the same as used in the performance evaluation: the radiological reference is used to train the system, and that system is also evaluated against the radiological reference, idem for the external reference. The focal lesion detector was developed externally with independent and unknown training data.

Component	Training	Testing
Lung segmentation	*D-lung*	*D-eval*
Clavicle segmentation	*D-clavicle*	*D-eval*
Shape analysis	*D-shape*	*D-eval*
Texture analysis*	*D-eval*	
Focal lesion detection†	N/A	*D-eval*
Score normalization*‡	*D-eval*	
Supervised combination*	*D-eval*	

N/A Not available
** Training and testing is performed in crossvalidation
† The focal lesion detection system was externally developed and there is no information on the used training sets
‡ Score normalization requires a training set but no label information

Table 6.4: Training and testing sets of the components and systems used in the paper. The left side of the table shows the datasets used for the systems: *D-lung* is disjoint from *D-eval*, and *D-shape* and *D-clavicle* are subsets of *D-lung*.

Dataset	Find & Treat	TB-NEAT
D-eval	200	200
D-lung	495	548
D-clavicle	250	548
D-shape	483	225

Table 6.5: Number of images in the training sets for the Find & Treat and TB-NEAT database.

6.B Results for all combination methods

This appendix provides results of all (supervised and unsupervised) combination methods. For supervised combination linear discriminant analysis (LDA)[54], k-nearest-neighbor (kNN)[54], GentleBoost[58], and a Random Forest[59] classifier were used. For kNN we used $k = 13$ (KNN13), the odd value closest to the heuristic of $k = \sqrt{N}$[54], where $N = 180$ is the number of samples in the training dataset in the crossvalidation set-up. GentleBoost used 50 regression stumps as weak classifiers (GB50), and the RandomForest classifier used 50 decision trees with a maximum tree depth of 7 (RF50). The number of stumps and trees are based on pilot experiments, in which performances were not found to vary substantially with different parameters. Both GentleBoost and RandomForest are known to be resilient to overtraining[58,59]. Table 6.6 shows the performances of the four supervised combination methods (LDA, KNN13, GB50, and RF50) and the three unsupervised rules (Sum, Maximum, Product). The best combination method, as measured with average performance across databases and references, is RF50.

Database	Find & Treat		TB-NEAT	
Evaluation reference	External	Radiological	External	Radiological
RF50	**0.868**	**0.847**	0.741	**0.899**
GB50	**0.882**	**0.834**	0.738	**0.892**
LDA	**0.867**	**0.854**	0.728	**0.883**
KNN13	**0.865**	**0.854**	0.729	**0.875**
Sum	0.814	0.791	**0.765**	**0.891**
Maximum	0.823	0.8	0.758	0.866
Product	0.794	0.77	0.707	**0.875**

Table 6.6: Combination of subsystems with four supervised classifiers and three unsupervised rules. Combinations performing better than the best individual subscore are indicated in **bold**. The best combination method for the external and radiological reference are underlined. The table is sorted on the average performance across datasets and references, highest first.

Diagnostic accuracy of the automated system in a high-risk screening setting

7

Abstract

Tuberculosis (TB) screening programs may be optimized by decreasing the number of chest radiographs that require interpretation by human experts. The object of this study was to evaluate the performance of a computer aided detection software system in triaging active TB images from other images within a high throughput digital mobile TB screening program.

A retrospective evaluation of the software was performed on a large database consisting of 38,961 radiographs of unique individuals collected by the screening program between 2005 and 2010. In that period, 87 participants were diagnosed with active pulmonary TB. The software, consisting of a novel combination of textural and focal abnormality detection systems, generated a TB likelihood score for each radiograph. This score was compared to a reference standard of notified active pulmonary TB using Receiver Operating Characteristic analysis.

At a sensitivity of 95%, specificity was 48% (95% CI 30-61%), negative predictive value was 99.98% (95% CI 99.97% to 99.99%). The area under the Receiver Operating Characteristic curve of the software system in discriminating active TB cases from other cases was 0.86 (95% CI 0.82 to 0.89).

This study is the first to evaluate automatic software for TB detection in a large screening database. The software can identify almost half of the normal images in a TB screening setting with excellent sensitivity and therefore has potential to be used for triage.

7.1 Introduction

Tuberculosis (TB) remains one of the world's major health concerns. In 2011 an estimated 8.7 million new cases and 1.4 million deaths were reported. The majority of the TB burden is located in Africa, followed by the Asian countries[1]. Although the overall incidence of TB in the Western World has been decreasing, an increase in TB rates has been reported in selected high-risk populations, especially in urban settings[4,5].

Despite the increased effort to develop new TB diagnostics[13,244], screening is still commonly performed using chest radiography, followed by sputum culture or smear microscopy[21]. Several early studies have reported limited specificity and variable levels of inter- and intra-reader agreement the in interpretation of chest radiographs (CXRs) for TB detection[22]. The use of modern digital radiography and standardized scoring may lead to improved sensitivity and reader agreement[16,20,42,245,246]. Digital chest radiography is a quick and reliable technique with low marginal and operational costs[34].

In currently established screening programs large volumes of CXRs, that require manual assessment, are acquired from a high-risk population. High costs for image interpretation by human readers and the lack of skilled readers limit the potential of these programs. We propose to improve the efficiency of screening for TB using digital radiography by introducing computer-aided detection (CAD) in the workflow. CAD has been successfully applied in many fields of medicine, most notably mass detection in breast cancer screening[247], nodule detection in lung cancer screening[248], polyp detection in colorectal cancer[249], and automatic detection of diabetic retinopathy[250]. CAD has been employed in several ways to improve screening programs, through assistance of radiologists[251] or by providing precise disease quantification[252]. Detailed overviews have been published previously[50,253].

In this study, for the first time the effect of applying CAD to triage active pulmonary TB cases is determined. In a screening context triaging means that an initial triage test is used to identify suspect cases that subsequently require further diagnostic tests. These diagnostic tests are characterized by a higher specificity than the triage tests but also by higher costs and a larger burden on the screening participants. Examples of triage tests include fecal blood testing in colorectal cancer screening[254] and prostate specific antigen measurement in prostate cancer screening[255].

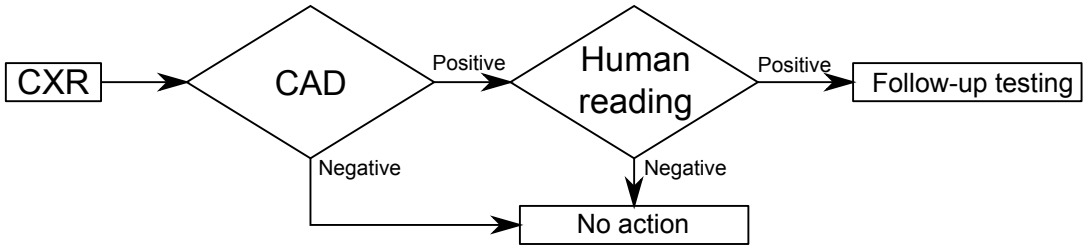


Figure 7.1: Proposed workflow with integration of computer aided detection (CAD) in chest radiographs (CXR) for TB detection as a first triage test before human reading.

The objective of this study is to describe a novel CAD system and to determine its benefit in a screening setting by evaluating its potential for triage. In the proposed workflow (Fig. 7.1) CXRs are analyzed automatically using computer software directly after acquisition. The output of the CAD system is a likelihood score for the image to contain TB related abnormalities. Cases with a score below a cut-off value are considered normal, and are removed from the screening process. Above the cut-off value, cases are considered suspect and are examined by a human reader. The ratio between cases judged normal by CAD and the total number of cases determines the reduction of reading burden.

7.2 Methods

The fully automatic CAD system for tuberculosis detection was applied to chest radiographs acquired by the Find&Treat screening program in the period 2005 - April 2010. The CAD score calculated for each radiograph was used to retrospectively determine the possibility of triaging active TB images with the goal of workload reduction.

7.2.1 Study population

A large image database consisting of 47,510 CXRs from 39,328 individuals was collected by the Find&Treat screening program in the period starting 1 April 2005 and ending 31 March 2010. This program runs in London, UK and screens a high-risk population consisting of homeless people, prisoners, and problem drug and alcohol users. All individuals attending the venue of the mobile X-ray unit were eligible for screening. CXRs were read directly after acquisition by a radiographer. The Find&Treat database has been described in detail previously[20]. Notified cases of active pulmonary TB were defined as cases for which a clinical decision was made to start TB treatment after radiological and symptom screening

(Class 3 according to the Diagnostic Standards and Classification of Tuberculosis[6]). In England 30% of notified pulmonary tuberculosis cases are not culture confirmed[256]. Since we were primarily concerned with sensitivity of CAD to enable triage, and because we wished to train the system to also be able to identify early changes in paucibacillary cases we used the clinical decision to treat, rather than culture as our comparator. For all participants with repeat radiographs only the most recent radiograph was used. Identification of repeat radiographs was based on a patient ID entered at the mobile unit. When a radiograph was linked to an active TB case that radiograph instead of the latest one was used. The other not active TB cases were either radiographically normal or contained abnormalities other than TB.

This study is a retrospective evaluation of an existing health care service using no identifiable patient information or possibility of deductive disclosure. The work had no impact on clinical management of the participants. According to the guidelines of University College London (where authors RWA, IA, ACH and AS are based) this exempts the study from review by a Research Ethics Committee[257].

7.2.2 Computer aided detection (CAD) of tuberculosis in chest radiographs

The proposed CAD system combines two systems, namely a texture analysis system and a focal lesion detection system, to analyze the CXR. The texture analysis system was previously developed by our group[147,148], whereas the focal lesion detection system is a commercially available software system. A key technique of the proposed CAD system is supervised classification, in which a statistical classifier is learned using manually labeled examples of object (pixels, image) in the training phase. In the test phase objects unseen in the training phase are assigned a probability of belonging to one of the learned labels.

Texture analysis

The texture analysis system consists of three components: segmentation of unobscured lung fields, detection of foreign objects and abnormality detection. The components use supervised pixel classification to perform their task. A large set of manually labeled pixels are used to train the classifier and, after classification, each pixel in a test image is assigned a label probability. The three components operate on images resampled to a fixed width of 1024 pixels:

1. Automatic segmentation of the unobscured lung fields. This component

uses a supervised pixel classification and a shape modeling method to identify the unobscured lung fields[76]. The lung field segmentation was used to restrict further analysis to this area and to provide anatomical context for the other components.

2. Automatic detection and removal of foreign objects. This component detects foreign objects (extracorporeal objects such as zippers, bra, clips , and buttons) using supervised pixel classification. Subsequently they were removed using a digital inpainting technique, which replaces each affected pixel by a similar pixel from the surrounding (unaffected) lung parenchyma[198].

3. Automatic detection of textural abnormalities. Textural abnormalities are broadly defined as non-focal changes in the visual properties of the lung tissue as a result of pathology. This component uses supervised pixel classification based on texture analysis to identify texture abnormalities[111,148]. The output of the classification is a heat map that indicates the probability that a pixel belongs to an abnormal region. The heat map was then summarized into one texture score (*texture score*) by computing the 95% quantile of its cumulative distribution. This texture score robustly measures the extent and the severity of affected lung[149].

Focal lesion detection

Focal lesions were automatically detected with commercially available software for nodule detection (ClearRead+Detect v5.2 ; Riverain Technologies, Miamisburg, Ohio). The software outputs for each image a list of suspicious locations with a likelihood score. The total load of focal lesions was summarized into one focal score (*focal score*) by adding the likelihood scores of all suspicious locations in the image detected by the software.

Combination of texture analysis and focal lesion detection

The textural analysis and focal lesion detection systems detect different types of abnormalities and one system may work better for an individual image than the other. It is well known that a combination of complementary systems can lead to an improvement in classification performance[62,234]. Therefore, the texture and focal scores were combined into one TB score. Combination was performed by a weighted average of the two scores: TB score = $a \cdot \text{texture score} + b \cdot \text{focal score}$.

This TB score indicates the overall TB suspiciousness of the image and was used to evaluate the performance of the system. The optimal values of the weights a and b were determined using linear discriminant analysis. Detailed descriptions of the software components are available in Appendix 7.A.

7.2.3 Procedures

Evaluation procedure

In this retrospective study, TB scores were generated for all CXRs in the dataset. Because of the limited number of available positive cases and to obtain unbiased performance estimates, the evaluation was performed using a five-fold cross validation procedure. In this procedure the available data was divided in five groups: in each fold four of the groups were used to train the system and generate scores for the remaining group. The final performance estimate is calculated as the average of the results obtained in each fold. The training of the texture analysis system, as well as the estimation of the optimal combination weights (a and b), was carried out using examples of active TB and normal images. Abnormal, but not active TB, cases were excluded from training but were classified in the testing phase. The fully automatic unobscured lung field and foreign object segmentation was separately trained with 495 cases, which were not used in the remainder of the study. Details are given in Appendix 7.A.

Annotation procedure

Annotated training examples needed for labels in pixel classification were created by the first author (LH), a Chest Radiograph and Recording System (CRRS)[42] certified reader, under supervision of an experienced chest radiologist. Outlines of textural abnormalities were created in active TB images. Outlines of unobscured lung fields and foreign objects were created in a sample of 495 images.

7.2.4 Outcome parameters and data analysis

The TB score was used to perform Receiver Operating Characteristic (ROC) analysis of CAD compared with the active TB reference. CAD's potential to triage was measured using the Negative Predictive Value (NPV) and specificity at a cut-off point of 95% sensitivity. Overall performance of the CAD system was measured using the Area under the ROC (AUC). AUC values were computed from ROC curves directly, i.e. no curve fitting procedures were employed. NPV, specificity, and AUC values were calculated with 95% confidence intervals using

a non-parametric bootstrap method[69]. In this method, a statistical distribution of ROC curves is constructed from 1,000 bootstrapped samples of the TB scores computed for the full dataset. Outcome parameters were calculated using the statistical software package R (R: A Language and Environment for Statistical Computing v2.13.1; R Foundation for Statistical Computing, Vienna, Austria).

7.3 Results

7.3.1 Patient population

Before removal of repeat scans, the database contained 47,510 CXRs. More men (36,948; 77.8%) than women (10,524; 22.2%) were screened. The highest number of CXRs were performed in 2006 (13,753; 28.9%) and the least in 2009 (7,866; 16.6%). Almost half of the screened persons were homeless (19,801; 41.6%); the remaining population consisted of prisoners, problem drug users, asylum seekers and other groups[5].

The radiographers identified 414 cases of suspected, yet unconfirmed, active pulmonary TB by CXR. There were also 458 cases of suspected old TB, 1,085 abnormal CXRs that were not referred and 124 abnormalities that were referred for further investigation. After clinical follow-up, 91 cases were confirmed as active TB cases. On the CXR reading, 86 of these 91 confirmed active TB cases had been identified as TB suspect, four cases were read as suspected old TB, and one case was referred for another suspicion than TB.

After removal of 8,282 duplicate CXRs, 39,328 CXRs of unique patients remained. Of these 367 could not be processed due to corruption of the CXR image data. After deduplication 87 active TB cases remained.

CAD scores were computed for 38,961 unique patients: 87 active TB cases and 38,874 other cases, of which 37,288 were normal and 1,586 abnormal, but not active TB cases. The NPV was 99.98% (95% CI 99.97% to 99.99%). Fig. 7.2 shows the ROC curve for the discrimination between active TB and other cases. The area under the curve was 0.86 (95% CI 0.82 to 0.89). At the cut-off point for triage the specificity is 48% (95% CI 30% to 61%) with 95% sensitivity. At this point 18,687 of 38,874 other cases and 83 of 87 active TB cases were correctly identified. Detecting 100% of the active TB cases reduced specificity to 26% (95% CI 25% to 48%).

In a representative sample of 1,000 images we found that approximately 20% of the CXRs in the database contained foreign objects. Fig. 7.3 shows three exam-

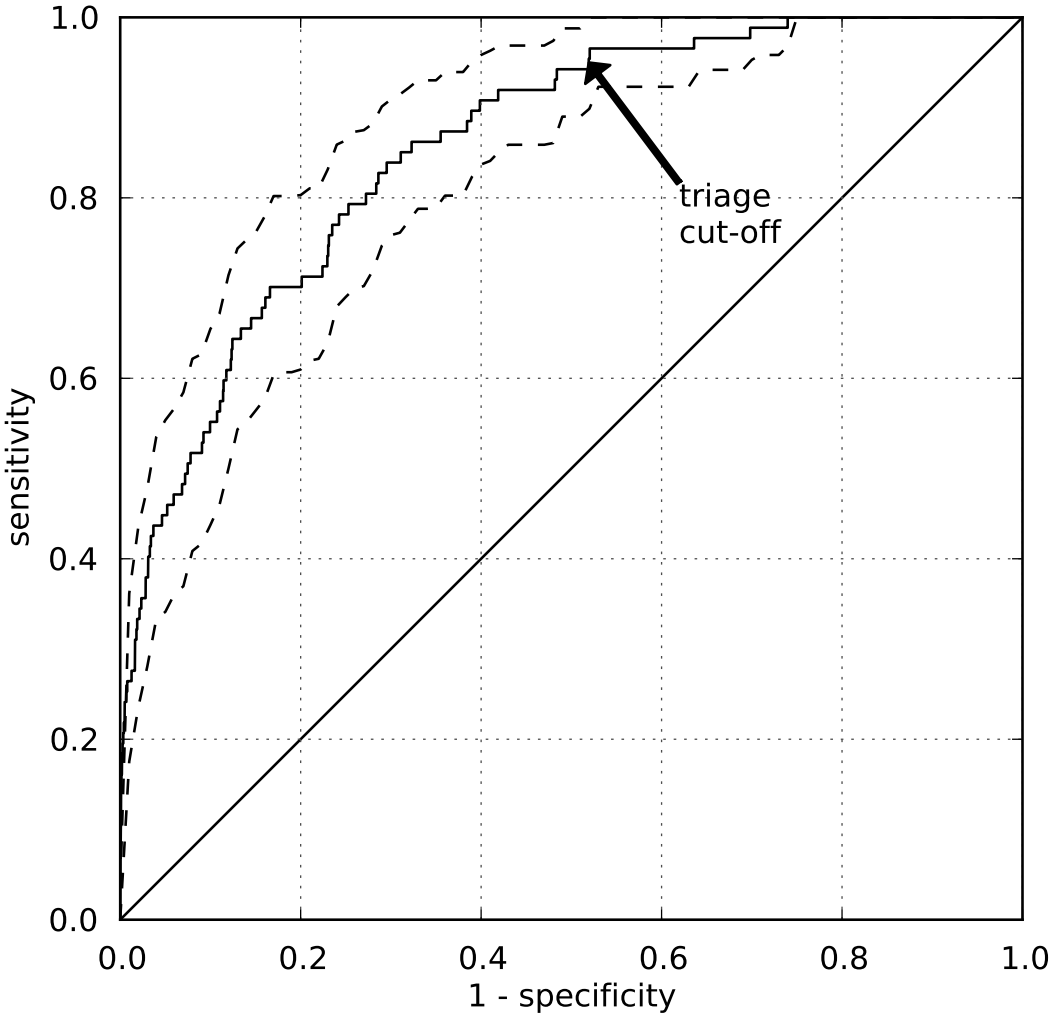


Figure 7.2: ROC analysis of CAD performance in discriminating between active TB and non active TB images. Dashed lines indicate the 95% confidence interval. The triage cut-off point at 95% sensitivity and 48% specificity is indicated by the arrow. Cases to the left of the cut-off point are referred for human reading, cases to the right are excluded from further analysis.

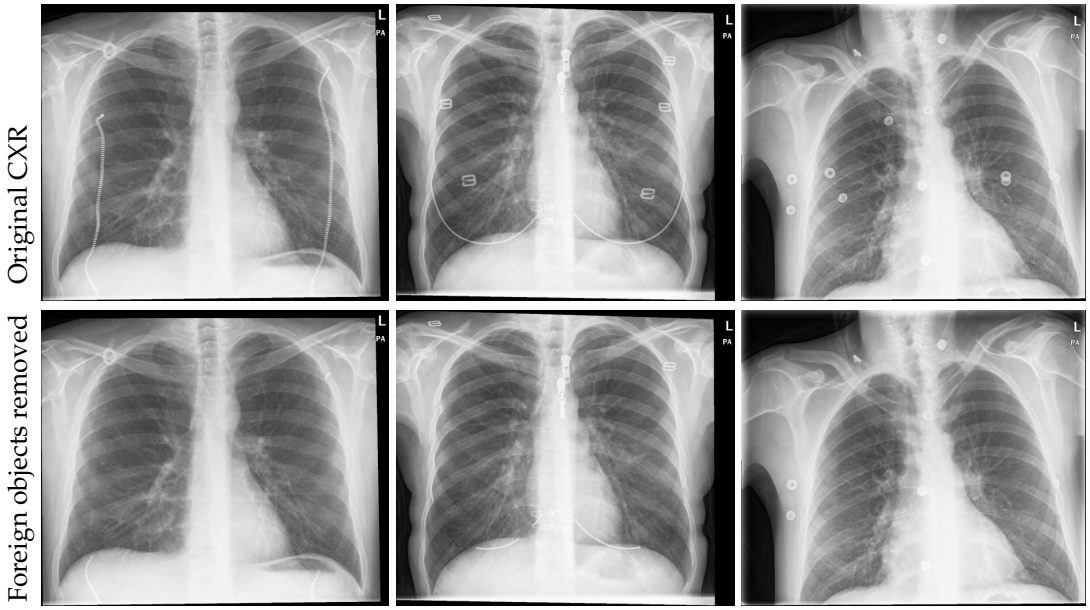


Figure 7.3: Examples of cases where foreign objects were removed using an automatic algorithm. The removal improves computation of the TB score by reduction of false positives responses.

ples of CXRs where the foreign objects were successfully detected and removed. The removal of foreign objects reduces the number of false positive detections of textural abnormalities in the lung regions. Fig. 7.4 shows results of the CAD system for a selected number of cases. The result of texture analysis and focal lesion detection is shown as a coded color map ranging from low probability to high probability through the colors blue-green-yellow-orange-red. The TB score reported per case in Fig. 7.4 is normalized to the fraction of all normal images that received a lower score (false positive fraction). The triage cut-off value corresponds to a normalized score of 52%.

7.3.2 Analysis of missed cases

At the cut-off point of 95% sensitivity, 4 of 87 active TB cases (5%) were missed by the software. A qualitative analysis of the cases was performed to determine the reason for their low TB score (Table 7.1). For this analysis, cases were categorized as containing predominantly parenchymal abnormalities (further divided in the groups very subtle, subtle, moderately clear, or clear), focal lesions, or lymphadenopathy. The cases reported as false negative by CAD were categorized as ”very subtle parenchymal abnormality” (1 case), ”focal lesions” (1 case), and ”lymphadenopathy” (2 cases) (Table 7.1; cut-off point = 95%/48%). Further

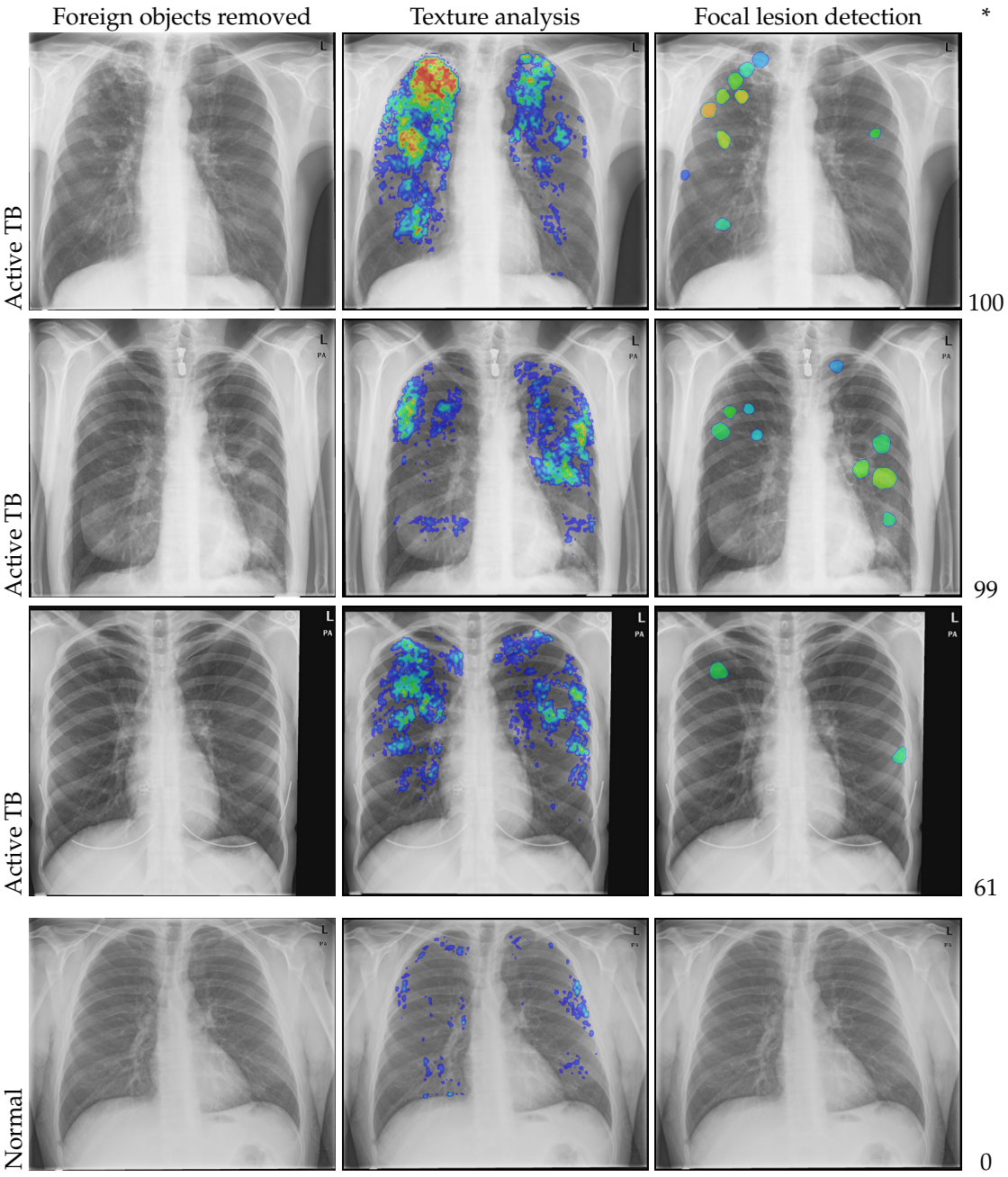


Figure 7.4: Examples of CAD analysis. First column: chest radiograph with foreign objects removed. Second and third column: color coded map of texture analysis and focal lesion detection, respectively. Colors indicate low to high abnormality suspiciousness in the order blue-green-yellow-orange-red. The last column (*) indicates the normalized TB score, equal to the percentage of normal cases receiving a lower score, ranging from 0 (normal) to 100 (abnormal).

Type of abnormality	Subtlety	Active TB cases missed Cut-off point (sensitivity/specificity) 95%/48%	 70%/80%
Parenchymal abnormalities	Very subtle	1	8
	Subtle	0	12
	Moderately clear	0	2
	Clear	0	0
Focal lesions		1	2
Lymphadenopathy		2	2

Table 7.1: Types of abnormalities in images missed at triage cut-off point (sensitivity/specificity = 95%/48%) and a high specificity cut-off point (sensitivity/specificity = 70%/80%). Analysis was based on a qualitative evaluation, see text for details.

increasing the potential to triage requires a lower number of false positive cases while maintaining the same level of sensitivity. At an increased specificity of 80%, sensitivity was 70% and 26 out of 87 cases were missed. Most of the extra missed cases correspond to very subtle and subtle parenchymal abnormalities (Table 1; cut-off point = 70%/80%).

7.4 Discussion

This retrospective evaluation of CAD for radiological TB screening on a large database found that the software can be used to exclude 48% of normal images from further reading, with a corresponding NPV of 99.9%, while maintaining a high sensitivity of 95%. A novel combination of textural and focal lesion detection was used. This study is also the first to evaluate a CAD system for TB in a large population in an operational screening setting. The CAD system was tested on an unselected set of images obtained in a Western screening setting that focuses on individuals at high-risk of TB. In such a setting, where the number of normal cases is typically much higher than abnormal cases, the use of a CAD system may result in a large reduction of the workload for human readers of CXRs, thus increasing the cost-effectiveness of screening programs[258].

The required sensitivity of a test in a TB screening program depends on the operational requirements of the setting in which it is employed. For TB prevalence surveys, the WHO handbook recommends over-reading to reduce the chance of missing cases[259]. Various performance measures of CXR for TB detection are reported in literature[16–19,21,246,260,261]: sensitivities range from 25%[17] to 95%[16,19] and specificities from 53%[18] to 99%[17]. A sensitivity of 95% was chosen for this study, equal to the highest reported value in literature, resulting in a specificity of 48%.

An important advantage of computerized reading is that all images are processed in a standardized, objective, and repeatable way. This facilitates the integration of CXR in a standardized screening protocol. Had the CAD system been prospectively used to triage active TB cases on the full Find&Treat database consisting of 47,510 CXR, a reduction of 48.1% (22,843) of cases referred to the human reader would have been achieved. A second advantage is that, unlike human readers who typically provide binary scores, CAD produces a continuous score which indicates the probability of the image corresponding to an active TB case. This continuous score can be used to set a specific cut-off point for different settings, in order to meet particular operational requirements.

In high-burden low resource countries, where the availability of skilled CXR readers is limited, CAD can be employed as the first and only reader. A cut-off point at high specificity but relatively low sensitivity can be used to reduce the number of normal cases that receive a follow up test to make the final diagnosis. While some active TB cases may be missed in this way, costs are lowered substantially as the total number of cases that require further testing is reduced. The use of CAD can make the implementation of national prevalence surveys more efficient. These surveys are recommended by the WHO to measure the TB burden and impact of TB control programs[262]. As they need to screen a large part of the population, reduced costs and a high throughput are highly advantageous. In the Western world, where radiological TB screening is mainly employed for high-risk and entrance screening, CAD can be used as a first reader and cases marked as suspicious by CAD can be read again by a human reader. The fractional reduction in workload in this scenario is almost directly proportional to the specificity as the percentage of active TB cases is generally small. Alternatively, CAD could be employed as second reader for quality assurance, in for example pre-entry screening programs.

Most systems for automatic analysis of CXRs have focused on single tasks, such as the detection of nodules[97] or general interstitial abnormalities[74]. The first paper directly addressing the topic of TB detection in CXRs was published by van Ginneken et al.[111]. In that study, texture analysis was employed to classify CXRs, acquired in a TB screening program for political asylum seekers, as normal or abnormal. Arzhaeva et al. reported on a different automatic system which classifies CXRs as normal or suspect for TB based on their global appearance[233]. Recently, Hogeweg et al. reported on an improved TB detection system, evaluated on a database of radiographs acquired in an African setting for TB suspect

screening, showing an improvement in performance by combination of local and global aspects of the radiograph[148]. This paper is the first to combine systems for two types of lesions, textural and focal, which both occur frequently in active TB cases.

The reference for the CAD evaluation was determined by active TB cases diagnosed by follow-up, that were all initially reported as radiologically abnormal. Additionally, at the time of the decision to refer clinical symptom information was available. In this study CAD only had access to the CXR. In a qualitative review of the four cases missed by CAD it was determined that three patients had one or more TB compatible symptoms at time of screening. Especially when operating at higher specificities, the presence of symptoms might increase the chance of referring CXRs with subtle abnormalities. With respect to human reading with knowledge of clinical information this can lead to an underestimation of CAD performance. Further research could investigate the combination of CAD with symptom information, although the latter is not always collected consistently in screening programs.

From the analysis of missed cases it follows that CAD had a high performance in detecting cases with textural parenchymal abnormalities. The majority of cases missed by CAD contained (very) subtle abnormalities or lymphadenopathy. Given the large variation of appearance of chest x-rays and the small number of confirmed active TB cases that were available to train the CAD system we expect that a larger dataset will improve the detection of subtle abnormalities. Lympadenopathy has a different appearance and location than other types of abnormalities. We expect for this category that, besides increasing the number of training examples, a dedicated subsystem focused at detecting changes in the hilar region could further improve performance.

7.5 Conclusion

We found that CAD can identify a large proportion of normal images in a TB screening setting at high sensitivity and has potential to be used for triage. This could increase cost-effectiveness of radiographic TB screening. Future work should focus on improving the CAD system to detect more types of abnormalities, further improvement of specificity, and prospective evaluation within screening programs.

Acknowledgments

This study was supported by the European and Developing Countries Clinical Trials Partnership (EDCTP), the "Evaluation of multiple novel and emerging technologies for TB diagnosis, in smear-negative and HIV-infected persons, in high burden countries" (TB-NEAT) project.

We would like to acknowledge the work of Jane Knight and Diana Taubman, the two reporting radiographers on the mobile X-ray unit in London who collected all of the CXRs.

Appendix

7.A Methods supplement

Supervised pixel classification

Texture analysis is based on a series of components in which each pixel in the chest radiograph is classified as belonging to a certain structure. Instead of binary classification, where each pixel is assigned to either one or the other label, soft classification is used, where a probability of belonging to a label is computed. In this way, first the lungs were segmented, then foreign objects were segmented, and finally textural abnormalities in the lung were detected. Pixel classification is based on supervised pattern recognition[51], in which during the training phase a statistical model is computed to discriminate between two complementary labels of pixels: in-/outside lung, in-/outside foreign objects, and in-/outside textural abnormalities. The statistical model is computed from a labeled training set of pixels, each of which is represented by a feature vector, a multi-dimensional numerical description of its characteristics. Many statistical models have been described in the literature[51]. We used the GentleBoost classifier, which is able to deal with nonlinear data, has low computational requirements, and is robust to overtraining[58].

Automatic segmentation of the unobscured lung fields

Only the unobscured lung fields were segmented, which are defined as pixels in the chest radiograph where radiation passes through the lungs but not through the heart, diaphragm or mediastinum. Segmentation is based on pixel classification and shape modeling, the algorithm is described in full detail in (3). The most important steps are described here.

Features

Local characteristics of each pixel in the image were computed. Two types of features were calculated: texture features based on Gaussian derivatives and features derived from the Hessian matrix. These features were computed from images resampled to a width of 512 pixels. To capture local image structure[133] the output of Gaussian derivative filtered images of order 0 through 2 (L, L_x, L_y, L_{xx}, L_{xy}, L_{yy}), at scales 1, 2, 4, 8, 16 pixels were calculated. To detect the presence of line-like structures Hessian matrix derived features, calculated at the same scales as the derivatives, were used (5). From the two eigenvalues of the Hessian matrix λ_1, λ_2, $|\lambda_1| > |\lambda_2|$ two measures were derived: (I) $\sqrt{\lambda_1^2 - \lambda_2^2}$ to extract the lineness of the local image structure and (II) the largest absolute eigenvalue $|\lambda_1|$ to indicate the strength of the response. In addition two position features, the normalized x- and y-coordinate, were added. Each pixel was thus described by a feature vector of 43 elements.

Training phase

From a set of 495 training images, examples of pixels in- and outside manually outlined lung fields were sampled (a random selection of 0.05%). The training set consisted of 63,566 pixels: 41,193 outside the lung fields and 22,373 inside the lung fields. A GentleBoost classifier with 100 regression stumps as weak classifiers was trained and used to classify pixels in test images. The result is a map containing for each pixel the probability that it is a lung pixel.

Shape modeling

In some cases abnormalities close to the unobscured lung field boundary might be excluded from the lung segmentation because of their high density. Shape information was included to improve the segmentation in these cases by applying a Hybrid Active Shape Model Pixel Classification algorithm[76,187]. The intensity model of the active shape model was trained on the probability map provided by pixel classification, instead of on the original image. The shape model was derived from the same training images as used in the pixel classification stage. The fitting of the shape model to a test image consists of an iterative procedure in which information from the probability map and shape model are balanced. The result of this step is an outline of the lung fields.

Automatic detection and removal of foreign objects

Detection and removal of foreign objects is described in full detail in Hogeweg et al.[198] (Chapter 2).

Features

The features used to detect foreign objects were the same as those used for the lung segmentation plus an additional four anatomical context features that were derived from the lung segmentation. These are the x- and y-coordinates relative to the bounding box of the lung fields, the distance to the lung wall, and the distance to the centroid of the lungs.

Training phase

Examples of pixels inside and outside foreign objects were sampled from a set of training images in which foreign objects were manually delineated. The training set consisted of 101 images containing one or more foreign objects selected from the same data set used to train the lung segmentation. In total 59,887 pixels were used, of which 49,268 were normal pixels and 10,619 were sampled inside foreign objects.

Object detection and removal

A GentleBoost classifier was trained with 1,000 regression stumps as weak classifiers and used to classify pixels inside the unobscured lung fields in test images. The resulting probability map was thresholded to segment foreign objects. Pixels inside segmented objects were replaced with similar looking pixels from the unaffected lung tissue using an inpainting technique[123]. Similarity was determined by comparing neighborhoods of pixels defined by a square patch of 11×11 pixels. Similar pixels were searched for in a region around the foreign object using a fast procedure based on approximate nearest neighbor search[136]. The removal of foreign objects ensures that the local statistical properties of the image, as measured by texture features, are not disturbed.

Automatic detection of textural abnormalities

The texture of an image refers to the appearance and structure of a small region. Textural abnormalities are one of the most common pathological changes caused by TB in chest radiographs. They can be the result of inflammatory processes in

the lung parenchyma or in the pleural space. The presence of textural abnormalities is determined by comparing the pixel density distributions of normal and abnormal regions.

Features

The texture analysis system provides for each pixel inside the lung field the probability of belonging to an abnormal region. Small circular image patches (radius = 32 pixels) were sampled every 8 pixels from radiographs resampled to a width of 1024 pixels, a higher resolution then used for segmentation in order to measure fine changes in the lung texture. Gaussian derivative filtered images were created at scales 1, 2, 4, and 8 pixels with derivatives of order 0 through 2 (L, L_x, L_y, L_{xx}, L_{xy}, L_{yy}), giving 24 filtered images. The features calculated for each patch were the first four moments (mean, standard deviation, skew, and kurtosis) of the intensity distributions of these filtered images and the original image. In total 100 texture features were computed. To add spatial information the same four anatomical context features as used in foreign object segmentation were added to the feature set.

Training phase

Textural abnormalities were outlined in the 87 active TB cases. The outlining was based on three abnormality patterns described in the Chest Radiograph Recording System[42]: small opacities, large opacities, and pleural fluid. From these outlines examples of abnormal pixels were selected. Examples of normal pixels were sampled from the set of other cases. A GentleBoost classifier with 100 regression stumps as weak classifiers was trained with approximately 350,000 pixels in each cross validation fold. The ratio between normal and abnormal samples was set to 30, which is approximately the average ratio between normal and pathological lung area in the chest radiographs.

Texture score

Each pixel inside the unobscured lung fields in a test image is classified, giving a probability that the pixel is part of textural abnormality. To determine the probability of the whole image being abnormal, all the individual pixel probabilities are summarized into one score. This score is computed by determining the 95% quantile of the cumulative distribution of pixel probabilities[149]. This texture score measures both the extent and severity of the affected lung region.

Focal lesion detection

Focal lesions were automatically detected with commercially available software for nodule detection (ClearRead+Detect v5.2 ; Riverain Technologies, Miamisburg, Ohio). The software outputs for each image a list of suspicious locations with a likelihood score. The total load of focal lesions was summarized into one focal score (focal score) by adding the likelihood scores of all suspicious locations in the image detected by the software.

Combination

A novel contribution of this paper is the combination of the output of texture analysis and a focal lesion detector. The idea behind this is that one system, as it is focused on one kind of abnormality, will not be able to perform well in all images. It is well known that a combination of complementary systems can lead to an improvement in classification performance[62,234]. Therefore the texture and focal score were combined into one TB score. The TB score indicates the overall TB suspiciousness of the image and was used to evaluate the performance of the system. Combination was performed by a weighted average of the two scores: TB score = $a \cdot \text{texture score} + b \cdot \text{focal score}$. The optimal values of the weights a and b were determined using linear discriminant analysis (LDA)[54]. Similar to pixel classification the combination requires a training phase, in this case with labeled training images. The image labels were set by the reference standard as active TB or other.

Summary and discussion

Summary and discussion

This final chapter discusses the main findings and conclusions of this thesis. The thesis presented several aspects of a CAD system for TB detection in chest radiographs, ranging from preprocessing, segmentation and feature design to a description and evaluation of a complete CAD system. The chapter starts with a summary of the main findings and its conclusions. In the general discussion overall conclusions and directions for future research and applications of the system are given.

Summary

Chapter 2 describes a method to detect, segment, and remove foreign objects from chest radiographs. Foreign objects regularly occur in large-scale screening settings where speed and patient comfort is important and people have therefore not been instructed to remove all their clothing. The method starts with a supervised classification of all pixels into foreign objects or normal lung tissue. Pixels classified as foreign objects were then grouped into objects and removed from the image with the goal of restoring the normal appearance of the lung parenchyma. Removal was performed by replacing each pixel in a foreign object with a similar looking pixel from the surrounding unaffected region. Similarity between pixels was determined by comparing the neighborhood of the pixel in a square patch. The effectiveness of the method was demonstrated in a texture analysis experiment concerning TB detection. The number of false-positive responses in the texture probability map was reduced by object removal. In this work mostly high density (metal) objects were removed. The method could be further extended by also detecting and removing foreign objects with a lower density.

Chapter 3 describes a method to segment the clavicles in chest radiographs. The clavicles are located in the upper part of the lungs, in the same area where TB appears most frequently. Together with other structures in this area the clavicles form a complicated pattern, which hampers analysis by humans and computers alike. In order to address the problem of segmenting the clavicles in this complex area a combination of several previously developed state-of-the-art techniques was used. First the location and an initial segmentation of the clavicle were found using pixel classification. A shape model was fitted to the initial segmentation to provide a plausible shape for the clavicle. Then the borders and the head of the clavicle were detected using specialized pixel classifiers. The information from

the pixel classifiers and the shape model were combined in a cost function. An optimal, precise outline of the clavicle was determined in this cost function by means of dynamic programming. The method was evaluated against a reference standard set by an human reference observer. Results showed an improvement compared to the previous best method and approached the performance of an independent human observer. This chapter showed that a combination of several methods can be used to improve performance. The clavicle segmentation was used in Chapters 4 and 6 to show the merit of suppressing them and using them as basis for a context feature, respectively.

In Chapter 4 a method was developed to suppress elongated structures, such as ribs and clavicles, in translucent images, like chest radiographs. Under translucent conditions it is possible to decompose projection images into their superimposed components. The problem was formulated by following blind source separation theory, where the goal is to isolate sources of interest in a mixture of unknown components. Elongated structures were modeled as a series of profiles perpendicular to the centerline of the structure. Blind source separation was performed using a PCA model of the profiles, assuming that the sources of interest correspond to the principal components. The model was then applied to the observed profiles to isolate the elongated structure from the image. Subtracting the isolated structure then removes it from the original image. The method was evaluated in four experiments on chest radiographs. In the first experiment ribs were removed from chest radiographs simulated from CT. The use of simulated chest radiographs provided an absolute reference standard. In the second experiment clavicles were suppressed in real chest radiographs and its effect on clavicle conspicuity was determined. In the third experiment the effect of clavicle suppression on the conspicuity of simulated nodules was determined. In the last experiment observers judged the automatic removal of catheters in chest radiographs. The results showed a marked reduction of the suppressed structures and suppression could be used in future research with the aim to improve the performance of the CAD system.

The lungs exhibit a large amount of symmetry in the chest radiograph. Deviations from normal symmetry might indicate the presence of abnormal structures, such as pathology caused by TB. In Chapter 5 we developed a method to measure symmetry both locally, in each point of the image, and globally, summarizing the overall symmetry of the image. Given a symmetry axis, which separates the image into a left and a right side, corresponding points were determined between

both sides. The correspondence between points was established by considering points that were most similar to each other in terms of a cost function consisting of the local intensity neighborhood of the point and its position. The value of the cost function between two corresponding points measures local asymmetry and the average of the local asymmetry over the whole image measures the global asymmetry. The method was evaluated in three experiments. In the first experiment, global asymmetry was used to classify chest radiographs as normal or abnormal. In the second and third experiments local asymmetry was shown to improve the results local texture analysis and also the contrast of nodules from a public database. The contribution of this method is twofold: (1) the unsupervised global measure provides a quick and robust way to detect the presence of relatively abnormal images, (2) the local asymmetry can be used to enhance texture analysis or other local analyses.

The presentation of TB on chest radiographs is diverse and different across populations, leading to a multitude of differently appearing abnormalities that have to be detected in order to reach a consistent performance with a CAD system. In Chapter 6 three different CAD systems were combined, detecting textural, focal and shape abnormalities. From these systems a number of subscores were extracted which describe different aspects of the chest radiograph. The subscores were collected into a feature vector, which was then used to classify an image as normal or TB. Classification was performed in a supervised manner, in which a classifier was constructed using a set of labeled training images. The method was evaluated on two data sets, one obtained from a high risk screening program in London, and the other obtained from TB suspects in Capetown. As a result of the differences in populations from which the two data sets were acquired, the relative frequency of different types of abnormalities is not the same. The individual systems, the combined system and the human observer were compared to a reference standard set by TB diagnosis and an experienced radiologist. The results showed that the combined system performed better than using a single subsystem, detecting either textural, focal or shape abnormalities. Also, supervised combination can be used to easily adapt the CAD system to different population or for different purposes, without having to retrain the individual systems.

In Chapter 7 a CAD system combining the detection of focal and textural abnormalities was evaluated in a large digital chest radiograph database acquired from a high risk screening program for TB in London, UK. Almost 40,000 chest radiographs from unique participants in a period of five years were retrospectively

analyzed by the CAD system. In the database 87 active TB cases were present. The CAD system consisted of a combination of focal and textural abnormality detection. The aim of the study was to demonstrate CAD's potential for triage, in other words selecting cases that need follow-up testing for TB. We found that the CAD system could eliminate almost half of the cases from further diagnostic tests while detecting 95% of the active TB cases. The other, suspect, cases can then be read by a human reader or subjected to a more specific diagnostic test for TB. Triage using CAD can reduce the workload in a screening program and therefore improve its efficiency.

General discussion

In this section first the main contributions of this thesis are discussed and then directions for future research are given.

A complete automated system for TB detection One of the main contributions of this thesis is the development and evaluation of a complete CAD system for TB detection in chest radiographs in Chapter 6 and a large scale evaluation in Chapter 7. A key principle in the design of the CAD system was that multiple subsystems are required in order to address the versatility of abnormalities that occur in TB. Not only does this versatility occur within a single patient population, but also across different populations. An automated system only detecting a single type of abnormality can have good performance in a particular population, but will not generalize well across populations, limiting its practical use. To obtain one decision from all the individual subsystems a generally applicable framework was presented to combine the results. Combination was performed by means of a pattern recognition approach with labeled training data. This strategy provides an efficient way to adapt the CAD system to different populations or for different tasks.

High-level combination of automated systems The medical image analysis researcher encounters a plethora of different approaches, methods, algorithms, and implementations. It is very difficult to decide a priori which approach is optimal for a particular task. In this thesis this problem was addressed by a combination of existing approaches. In Chapter 3 it was found that a combination of segmentation techniques for clavicle segmentation performed significantly better than previous methods. In Chapter 6 it was shown that a combination of sys-

tems detecting different abnormalities improved performance compared to the individual systems. The question is which systems to combine, how to combine them, and how to decide the contribution of each of the systems.

To perform well in combination, systems preferably make different kind of errors[60]. A considerable body of literature is concerned with ensembles of systems (classifiers) which fulfill these properties by design[58,59], but these systems all perform the same task. Instead, in this thesis combination of systems at a higher level, namely systems that address different subproblems of the overall task, was performed. High performing individual systems will likely contribute in combination, but not all systems are required to have very high performance. For example in a nodule detection task on CT it has been shown that the combination of the three worst performing systems had considerably better performance than the best system[61]. In another study on classifier combination it was found that even adding the worst performing systems contributed to the combination[263].

To combine systems their outputs have to be brought in a similar format, so that they can be combined in a uniform framework. For segmentation in Chapter 3 a cost function paradigm was used, in which at every location in the image evidence is built up about the presence or absence of the structure of interest. In this cost image the optimal combination of evidence was efficiently found using dynamic programming. In Chapter 6, which describes the full CAD system, the suspiciousness scores of the individual systems were combined into a feature vector. Adding label information to this feature vector reformulates combination as a pattern recognition problem.

The final consideration is how to weight each of the systems in the combination. In the segmentation framework this was done by greedy searching for the optimal weights for each of the components (systems) in the cost function. In the CAD framework weighting was determined in the training procedure of the classifier used for combination.

Removal of unwanted structures Pathology detection can be considered as the process of ignoring everything that is not relevant for the task. In this thesis, Chapters 4 and 2 address the removal of irrelevant structures. The problem of anatomical noise in the chest radiograph, which complicates human and computer analysis, is addressed by suppressing normal bony anatomy. Several previous studies have addressed this problem before[88,89,93], but the most promising ones[90,91] depend on training data acquired from dual energy radiographs, which

makes it difficult to apply such a system to new data. Instead the method in Chapter 4 describes an unsupervised technique to remove the ribs and clavicles, which showed promising results. Detecting relevance in unsupervised methods is a difficult problem though and we noticed some problems when a large number of other structures were overlapping. These problems were partially alleviated by detecting and removing outlying structures from the data, but this did not always succeed when the other overlapping structures were too prominently present.

Further improvement of the method for suppression of the bony structures could lead to a better detection of small or subtle abnormalities. These categories of abnormalities are, besides uncommon types of abnormalities, the most difficult for the current CAD system. The positive effect of bone suppression has already been demonstrated for the detection of nodules in chest radiographs[85,87], but not for textural abnormalities, which are the more common manifestation of TB. In Chapter 2 it was shown that the automatic detection and removal of foreign objects from the chest radiograph improved the detection of textural abnormalities. Foreign objects are easy to ignore as irrelevant for human readers, but should be addressed by the automated system to prevent false positives. We expect that both techniques will be integrated in future versions of the CAD system.

Unsupervised detection of pathology Chapter 5 on symmetry demonstrated that without any label information good image classification performance can be achieved. This result is explained by the self-normalizing property of asymmetry computation, i.e. pixels from the same image are compared to each other to determine the presence of abnormalities. This property avoids, amongst others, issues with differences in lung appearance between patients. In future research we are planning to extend the method to compare multiple CXRs of the same patient, for example to monitor treatment response[264]. On the other hand, the texture analysis method in this thesis requires a (large) set of labeled pixels from a training set. Pixel labels for texture analysis were determined by manual outlining of the abnormalities in the image. It is difficult to be completely sure that a pixel is normal or abnormal, due to the complex pattern induced by the superimposition of structures, while most supervised classification techniques assume that the provided label information is reliable. This problem was addressed before in a study by Arzhaeva et al.[233], in which the label information at the pixel level was dispensed with altogether. Instead, a dissimilarity-based approach was used, in

which image characteristics were compared to a set of labeled prototype images. Like the asymmetry measures, this method was shown to achieve similar results as a local pixel-based texture analysis, although it still requires image labels. Another way to classify images without local label information stems from the field of *content-based image retrieval* (CBIR). In these approaches an image is compared to a large set of labeled images and a selection is made of the most similar images. By counting the number of normal and abnormal ones among the similar images a probability of being abnormal can be assigned[224].

Directions for future research

How could the automated system be integrated into screening algorithms? A main goal of the CAD system described in this thesis is to provide an automatic screening tool for TB that can be included as part of a point of care test[12,265]. The question is how to integrate a CAD system in current screening practice such that its utility is maximal. For TB screening a number of strategies have been defined by the WHO. Examples are symptom screening followed by sputum testing or chest radiography followed by sputum testing[21,28]. We propose to either completely replace the human reading of radiographs or to reduce the number of cases that require human reading. Simulation studies should be performed to determine the effect of integrating CAD with these strategies on aspects such as detection rates, false positives, and costs[29,266]. In a study in Zambia* a prototype of the CAD system replaced human reading in order to select subjects that required GeneXpert testing. One of the reasons for this choice was the unavailability of a human reader at the moment of the study; such practical or logistic issues are common in the countries with the highest rates of TB. In these settings, we expect that CAD could contribute quickly to the diagnostic process.

Chapter 7 describes the evaluation of the CAD system in a setting, where the basic idea is to triage (filter out) cases that require human CXR reading. For such a first screening test it is important that a high level of sensitivity is achieved. Because screening programs are concerned with many more healthy subjects than ill subjects, the achieved specificity will largely determine how much cases could be excluded from human reading. Although the established opinion on screening tests is that they should have high sensitivity, this requirement can be lessened if screening is repeated after a certain amount of time[267]. Although for TB it is

*Unpublished results. The study was executed by the ZAMBART organization, which is one of the research partners in the CAD4TB project.

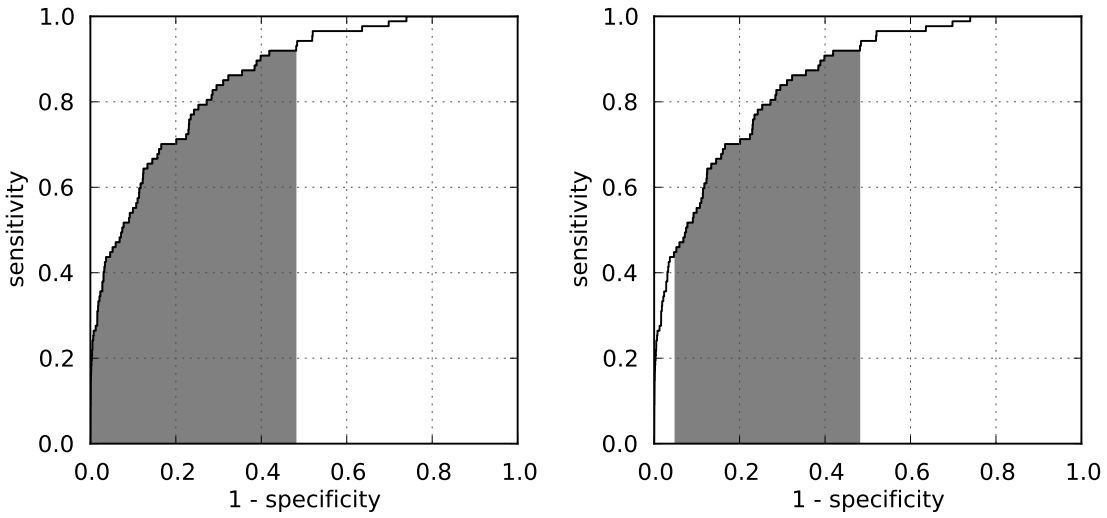


Figure 7.5: Referral of cases to human reader in a first reader (CAD)/second reader(human) paradigm based on threshold on TB score. The cases in the shaded part of the curve are referred (a) Referral of abnormal (most suspicious) cases. (b) Referral of uncertain cases. Cases with high certainty of being abnormal (leftmost part of the curve) and being normal (rightmost part of the curve) are not referred. Example from Chapter 7.

important that cases are detected early, to prevent infection of other people, a repeat screening strategy might be the best option when confirmatory diagnostic tests, such as sputum culture or GeneXpert, cannot be supplied to everyone.

How could automated scoring be integrated in the reading process? There are different options for applying the output of the automated system. For example, the automated TB score or visual indications of the location of abnormalities could be displayed to aid human readers. To improve screening efficiency a first reader/second reader paradigm is more appropriate; CAD as the first reader operates autonomously, the human reader is only called in for help when the software is uncertain about its decision. This strategy requires the determination of a threshold on the automated system's TB score and a decision on which images (abnormal, or only uncertain ones) to "refer" to the human reader. The automated TB score indicates the likelihood that the given image contains TB related abnormalities. Very high and low scores indicate that the system is certain about the label of being abnormal and normal, respectively. By defining performance targets – based on the expected number of normal and abnormal cases, but also on available human, material, or financial resources – a threshold on the TB score can

be determined to select cases that do not require human reading. In chapter 7, on triaging active TB cases, the decision was made to refer only abnormal cases for human reading (Fig 7.5(a)). Through the ROC curve, the required sensitivity determines the specificity and how many normal cases can be excluded from human reading. If the sensitivity is set below 100% (95% in the example) it leads to missing a small fraction of difficult cases. On the other hand, if it is demonstrated that CAD can identify a selection of TB cases with a high degree of certainty, then also these could be exempt from human reading (Fig 7.5(b)). This latter strategy is mainly of interest when the number of abnormal cases is relatively high.

Which types of abnormalities should receive more attention? Analysis of the results showed that two categories of abnormal images were difficult to classify correctly: images containing abnormalities that are not specifically addressed in the automated system and images containing subtle abnormalities. The first group contains abnormalities such as lymphadenopathy, pleural effusions, blunting of the costophrenic recess (CPR) and small focal lesions. The problem of detecting blunting of the CPR has been addressed before[243,268], and such a method could be added in the combined CAD system. To our knowledge no previous method has specifically addressed the detection of lymphadenopathy. It is difficult to collect a good reference dataset though, because human readers often disagree about the presence of lymphadenopathy in chest radiographs[246,264]. In such a situation it would be helpful to have CT images accompany the chest radiograph so that a reliable reference can be created, such as has been done by Arzhaeva et al.[147]. Unfortunately, CT images are not routinely acquired in TB screening or diagnosis in countries with the highest incidence and collecting a large dataset will be difficult. Small focal lesions (nodules) could be detected with a modified CAD system for larger nodules[97]. There also have been a number of previous studies concerned specifically with the automatic detection of pneumoconiosis[269,270]. Although this is a different disease, it also presents as multiple small nodules in chest radiographs.

The latter group of difficult abnormalities consists of subtle variations of any of the types of abnormalities. Besides the earlier discussed suppression of normal anatomy, improvements are expected by providing a larger number of examples in the training sets. Increasing the size of the training data will not always work though[271] and might result in prohibitively long computation times. Also, all this training data should then be manually annotated. These kinds of problems are addressed by the concept of *active learning*[272]; instead of indiscriminately

adding all available examples to the training set, only those who are interesting are added. The interestingness of an example is usually determined according to its position respective to the decision boundary; examples that are very close are difficult and likely interesting. With such a strategy, a large image database like the CAD4TB database*, could be quickly mined for interesting examples, without human readers having to examine or annotate them all.

What other information could be used to improve the CAD system? Compared to the radiologist, or other human reader, the current CAD system is at a disadvantage in a clinical situation, because it has no information about the person's age, sex, medical history, and current symptoms. We expect that adding these types of characteristics as features to the combined CAD system described in chapter 6 will further improve its performance. Not all these characteristics are available in every setting or they are not collected in a structured way. Preferably, the CAD system should then be able to deal with missing information when combining the information from CXR and other sources. Several methods have been described in literature that are concerned with classification in the presence of missing values[273].

Which additional steps are required to use the CAD system in an operational setting? The next step in the CAD4TB project is that of applying the CAD system in an operational setting. Besides the integration in the screening and reading process, as discussed in previous paragraphs, a number of image processing related issues have to be addressed. One of the problems we encountered in extending the use of the CAD system to new settings is that differences in scanner technology, acquisition protocol, and custom post-processing lead to a different appearance of the image, causing a reduction in classification performance of the automated system. This is a well known general effect in supervised learning and its solutions part of the field of *transfer learning*[274]. Besides trying to correct for this during the learning (training) phase, images can be normalized before further processing to reduce acquisition related differences[275]. This is one of the approaches currently followed in the CAD4TB project[276]. The topic is slightly touched on in Chapter 6, where differences in the proportional area of the lungs in the CXR were removed. A related issue concerns whether the image is suitable at all for classification by the system. Examples where this is not the case are children, failed acquisitions, severe incorrect positioning of the patient, or lateral chest radiographs. To ad-

*More than 100,000 cases are available currently, although a small fraction with complete diagnostic information.

dress this issue, pathology detection must be preceded by a quality analysis step, such as described in Niemeijer et al.[117]. We are currently developing such a system for the CAD4TB project.

How reliable are the reference standards? A crucial part of any automated method that claims to be of practical value is the comparison to a reference standard. It is not always possible to create a reliable reference standard. As mentioned before, the pixel labels that were used for texture analysis in this thesis are not fully reliable. Sputum culture, the recommended reference standard for TB and used in Chapter 6 and Chapter 7, is also not perfectly reliable as a result of a limited sensitivity[242]. Additionally, good laboratory facilities and procedures are required to prevent cross-contamination of samples which can reduce specificity. These conditions can not always be fulfilled in low resource and high-burden settings. In practice sputum culture is not required to make a diagnosis. Diagnosis can be based on clinical judgment weighing the available evidence[6], or retrospectively based on treatment response[277]. When the reference standard is unreliable, it is helpful to compare the method to multiple reference standards or to determine human performance with respect to the reference. This was done in Chapter 6 for the evaluation of the complete CAD system and in Chapter 3 on clavicle segmentation.

What is the influence of the particular datasets used for evaluation? Datasets used in CAD evaluation, and TB detection specifically, do not only differ in size but also in the characteristics of the subjects and images. Because of this results of different researchers are difficult to compare and can result in wildly varying performances using the same method. From a methodological point of view this issue was addressed in Chapter 6 by combining several methods so that the robustness of the system across data sets is improved. Ideally, a large number of datasets collected from different populations are used to evaluate automated systems and to get a fair estimate of their performance. Collecting data and the accompanying reference standard can be a demanding task, which is often a considerable part of the research itself. Therefore, researchers are inclined to guard their carefully collected data, leading each research group often to use its own private data set.

Can data collection be improved? An organizational solution to the problem of data collection is to encourage sharing data among research groups, so that the effort of collecting data is divided among the groups. This requires the infrastructure to share data, and possibly also results and methods, but also an incentive for the researches to participate. One incentive, a political one, is to force researchers

to share data if that research is funded by public money*. In The Netherlands there is a tendency to impose such a policy. Another, perhaps more noble, incentive is that sharing data and methods will further the field and generate new possibilities for research. The Diagnostic Image Analysis Group, where this thesis was written, provides a platform for sharing†. The database in Chapter 3 was made publicly available‡ and there are plans to share a larger amount of the CAD4TB database.

Can combination of systems be performed more intelligently? In the set-up presented in Chapter 6 systems are all computed at the same time and then combined in parallel. With a large number of systems limited computational resources might reduce the practical usability of the CAD system. An alternative is to use serial combination, where systems are computed one at a time. The next system is only computed when the previous system is uncertain about its output. An example is the pixel classification scheme in van Ginneken et al.[76], where pixels are first classified by a low resolution system, and only pixels with uncertain labels are reclassified by higher resolution system. In a serial set-up with systems detecting different types of abnormalities the order of execution is important. A good choice is to start with systems with low computational requirements, which can detect the more abnormal cases, for example the shape abnormality detection system in chapter 6. Systems that detect the most difficult types of abnormalities, and typically have higher computational requirements, are only executed in a minority of cases. Examples are the detection of lymphadenopathy or subtle nodules. Alternatively the optimal order could be learned from the training data. A practical CAD system will likely be a hybrid of serial and parallel combination.

What other tasks could be addressed with the automated system? The CAD system described in this thesis was designed with the task of detecting active TB in mind. With relatively few modifications it could be adapted to other types of tasks. An important task, which was not addressed in this thesis, is to discriminate radiological patterns corresponding to active TB cases, from patterns corresponding to other diseases. A first challenge would be to discriminate active TB from old (healed) TB in the chest radiograph. Radiological differences have been described between these two situations[8] and it might be possible to encode these differences as image features in the combination framework presented in Chapter 6.

*The research in this thesis was supported by a mix of public and private funding

†`http://comicframework.org/`

‡`http://crass.comicframework.org/`

In a similar way other diseases such as pneumonia and pneumoconiosis could be detected by retraining the combination rule specifically for these diseases. A general CAD system for chest radiographs would be aimed at discriminating between all kinds of different diseases. Such a system might contribute to determine the cause behind sputum culture negative, but radiologically positive cases. The conflicting information in these cases complicate the clinical decision process and led to reduced performance of both humans and the automated system in Chapter 6.

In this thesis the system is evaluated based on a single TB score, although different aspects of the radiograph are evaluated. One way in which these different aspects could be used is by summarizing them in a structured report that can aid human reading. Structured reading of the chest radiograph by humans is currently already done in the form of scoring systems[18,46,264,278]. Future research could aim at filling scoring systems automatically, and provide insight into how human and automatic reading compare. The summary of the different aspects could also be used to measure similarity between cases and retrieve similar looking cases in a CBIR approach. Such approaches are interesting in teaching applications, where users examining a CXR are provided with visually similar radiographs, of which diagnostic and other available clinical information is known.

Epilogue

From an outsider's perspective it might seem strange that developing a CAD system which works properly in most cases appears to be so difficult. Especially when recent successes in machine learning are considered, such as the victory over human contestants in the Jeopardy game by IBM's Watson and the autonomously driving car by Google. One reason for these projects being so successful is that they are backed by large companies which have the possibility to invest large amounts of financial resources. It could be that the problem of TB detection in chest radiographs could quickly achieve human expert performance once IBM, Google, or Microsoft take interest. The chance that they will be doing so is small though, because return on investment is limited or absent in the resource scarce settings in which TB is most problematic. Of course winning the Jeopardy game is not directly profitable for IBM either, but it is a showcase of their algorithms and computational facilities. The question is whether automatic TB detection is an attractive showcase.

One could also argue that the reading, automatic or human, of chest radio-

graphs is a very difficult task, which will not be quickly solved without advances in machine learning and pattern recognition. One such development is *deep learning*[279], which only recently attracted more attention because of its high computational requirements. Deep learning involves the extraction of high-level concepts from a large set of unlabeled data, and as such is a unsupervised learning strategy. Google recently demonstrated deep learning technology in which representations of high-level concepts, such as cats and humans, were automatically learned from a set of 10 million images extracted from YouTube videos using 16,000 computers crunching for three days[280]. The reason that such an approach might not be directly applicable to TB detection is not that its computational requirements are very high now; as long as Moore's law holds, which describes exponential increases in computation power for the same price, this same task could be performed by one computer in one day in 2040, and much earlier at other research labs around the world.

The reason that automatic TB detection, at the same level as a human expert, might not be quickly solved is a fundamental shortcoming in the field of artificial intelligence. This shortcoming is illustrated by the following approach to errors made by a supervised system, which is common to many researchers. As soon as the researcher finds, or thinks he has found, a pattern in the errors, he will modify his current method to incorporate more information; in this thesis often in the form of adding more systems. These modifications typically lead to new types of errors, although they might improve the overall performance of the system a little bit. This process of adding more and more information continues, but the CAD system still makes stupid errors; stupid in the sense that a human would never make such an error. This requirement to keep adding information is known as the context problem in artificial intelligence, and extensively discussed in a well-known critique of the field[281]. According to this critique the context problem cannot be solved because of the way (finite) numerical representations are required by computers. Another argument in the critique is that performing a complex task, such as reading chest radiographs, requires an intimate knowledge of the world in general, also known as common sense, and as a consequence thereof requires the intelligent apparatus to posses a body.

I would not like to conclude this thesis with ramblings about philosophical arguments concerning epistemology or how intelligence is defined, and I also would not like to argue that CAD systems should be robots or androids who also read poetry in order to understand chest radiographs. The message should be

that progress will be made by technological innovations and increasing computer power alike, but that before the system is as good as the best radiologist, it might already have contributed a bit in the regions of the world where people still die of TB every day.

Samenvatting

Hoofdstuk 2 beschrijft een methode voor het detecteren, segmenteren, en verwijderen van vreemde (lichaamsoneigen) objecten in thoraxfoto's. Vreemde objecten komen regelmatig voor in grote screeningsonderzoeken waar snelheid en patiënttevredenheid belangrijk zijn en deelnemers niet de instructie krijgen hun bovenkleding uit te doen. De eerste stap in de methode is een gesuperviseerde classificatie van alle pixels in de categorieën vreemd object of normaal long weefsel. Pixels geclassificeerd als vreemd object werden vervolgens gegroepeerd in objecten en digitaal verwijderd uit het beeld zodat de normale verschijning van het longweefsel in het beeld hersteld werd. Het verwijderen bestaat uit het vervangen van elke pixel in een vreemd object door een vergelijkbare pixel uit de omgeving, afkomstig uit normaal longweefsel. De effectiviteit van de methode werd aangetoond in een experiment waarin textuuranalyse werd gebruikt om TB te detecteren. Het aantal fout-positieve pixels in het TB waarschijnlijkheidsbeeld werd gereduceerd door het verwijderen van de vreemde objecten. In dit hoofdstuk werden voornamelijk vreemde objecten van hoge densititeit (bijvoorbeeld van metaal) verwerkt. De methode kan verder uitgebreid worden om ook objecten met een lage densiteit te detecteren en te verwijderen.

Hoofdstuk 3 beschrijft een methode voor het segmenteren van de sleutelbeenderen in thoraxfoto's. De sleutelbeenderen bevinden zich in de bovenste longvelden, hetzelfde gebied waar TB het meest voorkomt. De sleutelbeenderen vormen samen met andere structuren een ingewikkeld patroon, die analyse door mensen en computers bemoeilijkt. Om de sleutelbeenderen te segmenteren in dit complexe gebied is een methode ontwikkeld die een aantal eerder ontwikkelde *state-of-the-art* methodes combineert. Als eerste werd een initiële segmentatie van het sleutelbeen bepaald door middel van pixelclassificatie. Een vorm-gebaseerd model werd vervolgens toegepast op de initiële segmentatie om een aannemelijke vorm te selecteren. Vervolgens werden de randen en het mediale gedeelte van het sleutelbeen gedecteerd door middel van gespecialiseerde pixelclassificatie. De informatie van deze pixelclassificatie en van het vorm-gebaseerd model werden gecombineerd in een kostenfunctie. Een optimale en nauwkeurige bepaling van de rand van het sleutelbeen werd gevonden door middel van *dynamic programming*. De methode werd geëvalueerd door deze te vergelijken met een menselijke referentiestandaard. De resultaten toonde een verbetering ten opzichte van eerdere methodes en de nauwkeurigheid van een onafhankelijke menselijke lezer werd benaderd. Dit hoofdstuk toonde aan dat een combinatie van een aantal bestaande methodes gebruikt kan worden om de resultaten te verbeteren. De sleutelbeen-

segmentatie is gebruikt in hoofdstukken 4 en 6 om het effect van het verwijderen ervan en het gebruik als een context-eigenschap te bepalen.

Hoofdstuk 4 beschrijft een methode voor het onderdrukken van langwerpige structuren, zoals ribben en sleutelbeenderen, in doorschijnende beelden, zoals thoraxfoto's. Als aan de voorwaarde van doorschijnendheid is voldaan, dan is het mogelijk om projectiebeelden te ontbinden in hun overlappende componenten. Het probleem werd geformuleerd binnen de theorie van *blind source separation*, waarin het doel is om bepaalde relevante *sources* te isoleren uit een mengsel van onbekende componenten. Langwerpige structuren werden gemodeleerd als een reeks profielen die haaks staan op de middellijn van de structuur. Blind source separation werd uitgevoerd door middel van een PCA model van de profielen, met de aanname dat de relevante bronnen overeenkomen met de principale componenten. Dit model werd vervolgens toegepast op de profielen om de langwerpige structuur in het beeld te isoleren. De geïsoleerde structuur werd vervolgens uit het originele beeld verwijderd. De methode werd geëvalueerd in vier taken. In de eerste taak werden ribben verwijderd uit thoraxfoto's, die gesimuleerd waren op basis van CT beelden. Het gebruik van gesimuleerde beelden zorgde voor nauwkeurig referentiemateriaal. In de tweede taak werden sleutelbeenderen onderdrukt in echte thoraxfoto's en het effect daarvan op de zichtbaarheid van het sleutelbeen werd bepaald. In de derde taak werd het effect van sleutelbeenonderdrukking op de zichtbaarheid van gesimuleerde nodulen bepaald. In de laatste taak werd het automatisch verwijderen van katheters in thoraxfoto's beoordeeld door een aantal lezers. De resultaten toonde een duidelijke vermindering van de zichtbaarheid van de onderdrukte structuren. Deze methode kan in vervolgonderzoek gebruikt worden om de prestaties van het CAD systeem te verbeteren.

De longen vertonen een grote mate van symmetrie op de thoraxfoto. Afwijkingen van normale symmetrie kunnen duiden op de aanwezigheid van abnormale structuren, zoals pathologie als gevolg van TB. In Hoofdstuk 5 hebben we een methode ontwikkeld om symmetrie te meten: lokaal (op elk punt in het beeld) en globaal (de totale hoeveelheid symmetrie in het beeld). Met behulp van een symmetrie-as, die het beeld in een linker- en rechtervlak deelt, werden corresponderende punten aan beide kanten bepaald. De correspondentie tussen twee punten werd vastgesteld door voor elk punt een punt aan de andere kant te vinden, die het meest gelijkend is in termen van een kostenfunctie. Deze functie bestaat uit een combinatie van de lokale intensiteit en positie. De waarde van de

kostenfunctie tussen twee corresponderende punten bepaalt de lokale asymmetrie en het gemiddelde van alle lokale asymmetrie waarden bepaalt de globale asymmetrie. De methode werd geëvalueerd in drie experimenten. In het eerste experiment werd de globale asymmetrie gebruikt om thoraxfoto's als normaal of abnormaal te classificeren. In het tweede en derde experiment werd lokale asymmetrie gebruikt voor respectievelijk het verbeteren van de resultaten van lokale textuuranalyse en het verhogen van het contrast van nodulen afkomstig uit een publiek beschikbare database. De bijdrage van deze methode is tweeledig: (1) de niet-gesuperviseerde globale asymmetrie kan gebruikt worden als een snelle en robuuste methode om abnormale beelden te detecteren, (2) de lokale asymmetrie kan gebruikt worden voor het verbeteren van lokale analyses, zoals textuuranalyse.

De verschijning van TB op thoraxfoto's is divers en verschilt per populatie. Dit vereist het kunnen detecteren van verschillende typen afwijkingen om consistent goede prestaties te bereiken met een CAD systeem. In hoofdstuk 6 werden drie verschillende CAD systemen gecombineerd die respectievelijk texturele, focale, en vormafwijkingen detecteren. Elk systeem heeft als uitkomst een of meer subscores, die verschillende aspecten van de thoraxfoto beschrijven. De subscores werden gecombineerd in een vector van eigenschappen, die vervolgens gebruikt werd om een beeld te classificeren als normaal of als verdacht voor TB. Het classificeren werd gesuperviseerd uitgevoerd, door middel van een *classifier* die getraind werd met een database van gelabelde trainingsbeelden. De methode werd geëvalueerd op twee datasets, één afkomstig van een screeningprogramma voor hoog-risico groepen in Londen, en één afkomstig van deelnemers uit Kaapstad, die verdacht werden van TB. De relatieve frequentie van verschillende types abnormaliteiten verschilt in de twee datasets als gevolg van de verschillen in beide populaties. De drie subsystemen, het gecombineerde systeem en een onafhankelijke lezer werden vergeleken met twee referentiestandaarden: de één gebaseerd op TB diagnose en de ander afkomstig van een ervaren radioloog. De resultaten toonde aan dat het gecombineerde systeem betere prestaties heeft dan een enkel subsysteem, die óf texturele, óf focale, óf vormafwijkingen detecteert. Een dergelijk gecombineerd systeem, gebaseerd op een trainingsdatabase, kan gebruikt worden om het CAD systeem eenvoudig aan te passen voor gebruik in verschillende populaties of voor andere doeleinden.

Hoofdstuk 7 beschrijft een CAD systeem en de evaluatie in een grote database van digitale thoraxfoto's, die afkomstig waren uit een screeningprogramma voor

hoog-risico groepen in Londen, Verenigd Koninkrijk. Bijna 40,000 thoraxfoto's, afkomstig van individuele deelnemers, verzameld in een periode van 5 jaar, werden retrospectief geanalyseerd door het CAD systeem. De database bevatte 87 gevallen van actieve (besmettelijke) TB. Het CAD systeem bestond uit een combinatie van twee systemen: één voor focale afwijkingen en één voor texturele afwijkingen. Het doel van de studie was om de mogelijkheid te onderzoeken om CAD te gebruiken als *triage* systeem, dat wil zeggen: het selecteren van deelnemers die een vervolgtest behoeven. De resultaten toonden aan dat het CAD systeem bijna de helft van de deelnemers van vervolgonderzoek kan uitsluiten. In dit geval werd 95% van de actieve TB gevallen gedetecteert. De overgebleven verdachte gevallen worden dan gelezen door een menselijke lezer of krijgen een meer specifieke diagnostische TB test. Triage door middel van CAD kan de werklast in screeningprogramma's beperken en daarmee hun effectiviteit vergroten.

Publications

Papers in international journals

L. Hogeweg, C. I. Sánchez, P. A. de Jong, P. Maduskar, and B. van Ginneken. Clavicle segmentation in chest radiographs. *Medical Image Analysis*, 16(8):1490 – 1502, 2012.

L. Hogeweg, C.I. Sánchez, J. Melendez, P. Maduskar, A. Story, A. Hayward, and B van Ginneken. Foreign object detection and removal to improve automated analysis of chest radiographs. *Medical Physics*, 40(7):071901, 2013.

L. Hogeweg, C.I. Sánchez, and B. van Ginneken. Suppression of translucent elongated structures: applications in chest radiography. *IEEE, Transactions on Medical Imaging*, in press, 2013.

L. Hogeweg, C.I. Sánchez, R.W. Aldridge, A.C. Hayward, I. Abubakar, B. van Ginneken, and A. Story. Diagnostic accuracy of an automated system for detection of tuberculosis on chest radiographs: potential for triage in high-risk screening. *Submitted*, 2013.

L. Hogeweg, C.I. Sánchez, P. Maduskar, R.H.H.M. Philipsen, and B. van Ginneken. Fast and effective quantification of symmetry in medical images: application to chest radiography. *Submitted*, 2013.

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

P. A. de Jong, J. A. Achterberg, O. A. M. Kessels, B. van Ginneken, **L. Hogeweg**, F. J. Beek, and S. W. J. Terheggen-Lagro. Modified Chrispin-Norman chest radiography score for cystic fibrosis: observer agreement and correlation with lung function. *European Radiology*, 21:722-729, 2011.

Papers in conference proceedings

L. Hogeweg, C. Mol, P. A. de Jong, and B. van Ginneken. Rib suppression in chest radiographs to improve classification of textural abnormalities. In *Medical Imaging, volume 7624 of Proceedings of the SPIE*, pages 76240Y1–76240Y6, 2010.

L. Hogeweg, C. Mol, P. A. de Jong, R. Dawson, H. Ayles, and B. van Ginneken. Fusion of local and global detection systems to detect tuberculosis in chest ra-

diographs. In *Medical Image Computing and Computer-Assisted Intervention, volume 6363 of Lecture Notes in Computer Science*, pages 650–657, 2010.

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 *Medical Image Computing and Computer-Assisted Intervention, volume 5762 of Lecture Notes in Computer Science*, pages 724–731, 2009.

G.J.S. Litjens, **L. Hogeweg**, A.M.R. Schilham, P.A. de Jong, M.A. Viergever, and B. van Ginneken. Simulation of nodules and diffuse infiltrates in chest radiographs using CT templates. In *Medical Image Computing and Computer-Assisted Intervention, volume 6362 of Lecture Notes in Computer Science*, pages 396–403, 2010.

M.R.M. Samulski, P.R. Snoeren, B. Platel, B. van Ginneken, **L. Hogeweg**, C. Schaefer-Prokop, and N. Karssemeijer. Computer-Aided Detection as a Decision Assistant in Chest Radiography. In *Medical Imaging, volume 7966 of Proceedings of the SPIE*, pages 796614-1–796614-6, 2011.

P. Maduskar, **L. Hogeweg**, R.H.H.M. Philipsen, and B. van Ginneken. Automated localization of costophrenic recesses and costophrenic angle measurement on frontal chest radiographs. In *Medical Imaging, volume 8670 of Proceedings of the SPIE*, page 867038, 2013.

P. Maduskar, **L. Hogeweg**, R.H.H.M. Philipsen, S. Schalekamp, and B. van Ginneken. Improved texture analysis for automatic detection of tuberculosis (TB) on chest radiographs with bone suppression images. In *Medical Imaging, volume 8670 of Proceedings of the SPIE*, page 86700H, 2013.

R.H.H.M. Philipsen, P. Maduskar, **L. Hogeweg**, and B. van Ginneken. Normalization of chest radiographs. In *Medical Imaging, volume 8670 of Proceedings of the SPIE*, page 86700G, 2013.

Abstracts in conference proceedings

L. Hogeweg, C. Mol, P. A. de Jong, H. Ayles, R. Dawson, and B. van Ginneken. Evaluation of a computer aided detection system for tuberculosis on chest radiographs in a high-burden setting. In *Annual Meeting of the Radiological Society of North America*, 2010.

L. Hogeweg, A. Story, A. Hayward, R. Aldridge, I. Abubakar, P. Maduskar, and B. van Ginneken. Computer-aided detection of tuberculosis among high risk groups: potential for automated triage. In *Annual Meeting of the Radiological Society of North America*, 2011.

P. Maduskar, **L. Hogeweg**, H. Ayles, R. Dawson, P. A. de Jong, and B. van Ginneken. Automatic size measurement of cavities on chest radiographs using supervised learning and dynamic programming. In *Annual Meeting of the Radiological Society of North America*, 2011.

P. Maduskar, **L. Hogeweg**, H. Ayles, R. Dawson, P.A. de Jong, N. Karssemeijer, and B. van Ginneken. Cavity segmentation in chest radiographs. In *The Fourth International Workshop on Pulmonary Image Analysis*, 2011.

B. van Ginneken, **L. Hogeweg**, P. Maduskar, L. Peters-Bax, R. Dawson, K. Dheda, H. Ayles, J. Melendez, and C. I. Sánchez. Performance of inexperienced and experienced observers in detection of active tuberculosis on digital chest radiographs with and without the use of computer-aided diagnosis. In *Annual Meeting of the Radiological Society of North America*, 2012.

P. Maduskar, **L. Hogeweg**, H. Ayles, and B. van Ginneken. Performance evaluation of automatic chest radiograph reading for detection of tuberculosis (TB): a comparative study with clinical officers and certified readers on TB suspects in sub-Saharan africa. In *European Congress of Radiology*, 2013.

Bibliography

[1] World Health Organization. Global tuberculosis report 2012, 2012.

[2] Lawn S. and Zumla A. Tuberculosis. *Lancet*, 378:57–72, 2011.

[3] Zumla A., Raviglione M., Hafner R., and von Reyn C. F. Tuberculosis. *N Engl J Med*, 368: 745–755, 2013.

[4] de Vries G., van Hest R. A., and Richardus J. H. Impact of mobile radiographic screening on tuberculosis among drug users and homeless persons. *Am J Respir Crit Care Med*, 176: 201–207, 2007.

[5] Story A., Murad S., Roberts W., Verheyen M., Hayward A. C., and London Tuberculosis Nurses Network. Tuberculosis in London: the importance of homelessness, problem drug use and prison. *Thorax*, 62:667–671, 2007.

[6] Bass J., Farer L., P.C. H., Jacobs R., and Snider Jr D. American Thoracic Society. Diagnostic standards and classification of tuberculosis. *Am Rev Respir Dis*, 142(3):725–735, September 1990.

[7] Leung A. N. Pulmonary tuberculosis: the essentials. *Radiology*, 210:307–322, 1999.

[8] Woodring J., Vandiviere H., Fried A., Dillon M., Williams T., and Melvin I. Update: the radiographic features of pulmonary tuberculosis. *AJR Am J Roentgenol*, 146:497–506, March 1986.

[9] Steingart K. R., Henry M., Ng V., Hopewell P. C., Ramsay A., Cunningham J., Urbanczik R., Perkins M., Aziz M. A., Pai M., et al. Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. *Lancet Infect Dis*, 6(9):570, 2006.

[10] Siddiqi K., Lambert M.-L., Walley J., et al. Clinical diagnosis of smear-negative pulmonary tuberculosis in low-income countries: the current evidence. *Lancet Infect Dis*, 3:288, 2003.

[11] World Health Organization. Improving the diagnosis and treatment of smear-negative pulmonary and extrapulmonary tuberculosis among adults and adolescents: Recommendations for HIV-prevalent and resource-constrained settings, 2007.

[12] Dheda K., Ruhwald M., Theron G., Peter J., and Yam W. C. Point-of-care diagnosis of tuberculosis: past, present and future. *Respirology*, 18:217–232, 2013.

[13] Boehme C. C., Nicol M. P., Nabeta P., Michael J. S., Gotuzzo E., Tahirli R., Gler M. T., Blakemore R., Worodria W., Gray C., Huang L., Caceres T., Mehdiyev R., Raymond L., Whitelaw A., Sagadevan K., Alexander H., Albert H., Cobelens F., Cox H., Alland D., and Perkins M. D. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study. *Lancet*, 377:1495–1505, 2011.

[14] Sohn H., Minion J., Albert H., Dheda K., and Pai M. TB diagnostic tests: how do we figure out their costs? *Expert review of anti-infective therapy*, 7(6):723–733, 2009.

[15] Cleeff M. R. A. V., Kivihya-Ndugga L. E., Meme H., Odhiambo J. A., and Klatser P. R. The role and performance of chest X-ray for the diagnosis of tuberculosis: A cost-effectiveness analysis in Nairobi, Kenya. *BMC Infect Dis*, 5:111, 2005.

[16] den Boon S., White N. W., van Lill S. W. P., Borgdorff M. W., Verver S., Lombard C. J., Bateman E. D., Irusen E., Enarson D. A., and Beyers N. An evaluation of symptom and chest radiographic screening in tuberculosis prevalence surveys. *Int J Tuberc Lung Dis*, 10: 876–882, 2006.

[17] Lewis J. J., Charalambous S., Day J. H., Fielding K. L., Grant A. D., Hayes R. J., Corbett E. L., and Churchyard G. J. HIV infection does not affect active case finding of tuberculosis in South African gold miners. *Am J Respir Crit Care Med*, 180:1271–1278, 2009.

[18] Dawson R., Masuka P., Edwards D. J., Bateman E. D., Bekker L.-G., Wood R., and Lawn S. D. Chest radiograph reading and recording system: evaluation for tuberculosis screening in patients with advanced HIV. *Int J Tuberc Lung Dis*, 14:52–58, 2010.

[19] van 't Hoog A. H., Laserson K., Githui W., Meme H., Agaya J., Odeny L., Muchiri B., Marston B., Decock K., and Borgdorff M. High prevalence of pulmonary tuberculosis and inadequate case finding in rural western Kenya. *Am J Respir Crit Care Med*, 183(9):1245–1253, 2011.

[20] Story A., Aldridge R. W., Abubakar I., Stagg H. R., Lipman M., Watson J. M., and Hayward A. C. Active case finding for pulmonary tuberculosis using mobile digital chest radiography: an observational study. *Int J Tuberc Lung Dis*, 16:1461–1467, 2012.

[21] van't Hoog A. H., Meme H. K., Laserson K. F., Agaya J. A., Muchiri B. G., Githui W. A., Odeny L. O., Marston B. J., and Borgdorff M. W. Screening strategies for tuberculosis prevalence surveys: the value of chest radiography and symptoms. *PLoS One*, 7:e38691, 2012.

[22] Nyboe J. Results of the international study on x-ray classification. *Bulletin of the International Union Against Tuberculosis*, 41:115–124, 1968.

[23] National Collaborating Centre for Chronic Conditions (Great Britain) and Royal College of Physicians of London. Tuberculosis: clinical diagnosis and management of tuberculosis, and measures for its prevention and control. In *NICE Clinical Guidelines*. Royal College of Physicians, 2006.

[24] Jones T. F. and Schaffner W. Miniature chest radiograph screening for tuberculosis in jails: a cost-effectiveness analysis. *Am J Respir Crit Care Med*, 164:77–81, 2001.

[25] Coker R., Bell A., Pitman R., Hayward A., and Watson J. Screening programmes for tuberculosis in new entrants across europe. *Int J Tuberc Lung Dis*, 8(8):1022–1026, 2004.

[26] Golub J. E., Mohan C. I., Comstock G. W., and Chaisson R. E. Active case finding of tuberculosis: historical perspective and future prospects. *Int J Tuberc Lung Dis*, 9:1183–1203, 2005.

[27] Balabanova Y., Coker R., Fedorin I., Zakharova S., Plavinskij S., Krukov N., Atun R., and Drobniewski F. Variability in interpretation of chest radiographs among russian clinicians and implications for screening programmes: observational study. *British Medical Journal*, 331:379–382, 2005.

[28] World Health Organization. TB impact measurement: Policy and recommendations for how to assess the epidemiological burden of TB and the impact of TB control, 2009.

[29] Nishikiori N. and Van Weezenbeek C. Target prioritization and strategy selection for active case-finding of pulmonary tuberculosis: A tool to support country-level project planning. *BMC public health*, 13(1):97, 2013.

[30] Röntgen W. C. Über eine neue Art von Strahlen. *Sitzungsberichte der Physikalisch-Medicinisch Gesellschaft zu Würzburg*, pages 132–141, 1895.

[31] Mettler F. A., Bhargavan M., Faulkner K., Gilley D. B., Gray J. E., Ibbott G. S., Lipoti J. A., Mahesh M., McCrohan J. L., Stabin M. G., Thomadsen B. R., and Yoshizumi T. T. Radiologic and nuclear medicine studies in the United States and worldwide: frequency, radiation dose, and comparison with other radiation sources–1950-2007. *Radiology*, 253:520–531, 2009.

[32] Daffner R. H. *Clinical radiology, the essentials*. Williams & Wilkins, Baltimore, 2nd edition, 1999.

[33] Goodman L. *Felson's Principles of Chest Roentgenology*. Saunders, Philadelphia, 3rd edition, 2006.

[34] Zachary D., Schaap A., Muyoyeta M., Mulenga D., Brown J., and Ayles H. Changes in tuberculosis notifications and treatment delay in Zambia when introducing a digital x-ray service. *Public Health Action*, 2:56–60, 2012.

[35] Larson A., Lynch D., Zeligman B., Harlow C., Vanoni C., Thieme G., and Kilcoyne R. Accuracy of diagnosis of subtle chest disease and subtle fractures with a teleradiology system. *AJR Am J Roentgenol*, 170(1):19–22, 1998.

[36] Kotter E., Roesner A., Winterer J. T., Ghanem N., Einert A., Jaeger D., Uhrmeister P., and Langer M. Evaluation of lossy data compression of chest X-rays: a receiver operating characteristic study. *Invest Radiol*, 38:243–9, 2003.

[37] Shiraishi J., Katsuragawa S., Ikezoe J., Matsumoto T., Kobayashi T., Komatsu K., Matsui M., Fujita H., Kodera Y., and Doi K. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *AJR Am J Roentgenol*, 174:71–74, 2000.

[38] Agrons G. A., Markowitz R. I., and Kramer S. S. Pulmonary tuberculosis in children. *Semin Roentgenol*, 28:158 – 172, 1993.

[39] Chan C., Woo J., Or K., Chan R., and Cheung W. The effect of age on the presentation of patients with tuberculosis. *Tubercle and Lung Disease*, 76:290 – 294, 1995.

[40] Leung A., Müller N., Pineda P., and FitzGerald J. Primary tuberculosis in childhood: radiographic manifestations. *Radiology*, 182:87–91, 1992.

[41] Perlman D. C., el Sadr W. M., Nelson E. T., Matts J. P., Telzak E. E., Salomon N., Chirgwin K., and Hafner R. Variation of chest radiographic patterns in pulmonary tuberculosis by degree of human immunodeficiency virus-related immunosuppression. the Terry Beirn Community Programs for Clinical Research on AIDS (CPCRA). the AIDS Clinical Trials Group (ACTG). *Clin Infect Dis*, 25:242–246, 1997.

[42] Den Boon S., Bateman E. D., Enarson D. A., Borgdorff M., Verver S., Lombard C. J., Irusen

E., Beyers N., and White N. W. Development and evaluation of a new chest radiograph reading and recording system for epidemiological surveys of tuberculosis and lung disease. *Int J Tuberc Lung Dis*, 9:1088–1096, 2005.

[43] Miller Jr. W. T. Chest radiographic evaluation of diffuse infiltrative lung disease: review of a dying art. *Eur J Radiol*, 44:182–197, 2002.

[44] Nodine C. F., Kundel H. L., et al. Using eye movements to study visual search and to improve tumor detection. *Radiographics*, 7(6):1241–1250, 1987.

[45] Morgan R. Proficiency examination of physicians for classifying pneumoconiosis chest films. *AJR Am J Roentgenol*, 132(5):803–808, 1979.

[46] Pinto L. M., Pai M., Dheda K., Schwartzman K., Menzies D., and Steingart K. R. Scoring systems using chest radiographic features for the diagnosis of pulmonary tuberculosis in adults: a systematic review. *Eur Respir J*, 2012.

[47] Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph*, 31:198–211, 2007.

[48] Giger M. L., Karssemeijer N., and Armato S. G. Computer-aided diagnosis in medical imaging. *IEEE Trans Med Imaging*, 20:1205–1208, 2001.

[49] Lodwick G. S., Keats T. E., and Dorst J. P. The coding of Roentgen images for computer analysis as applied to lung cancer. *Radiology*, 81:185–200, 1963.

[50] van Ginneken B., Schaefer-Prokop C. M., and Prokop M. Computer-aided diagnosis: How to move from the laboratory to the clinic. *Radiology*, 261:719–732, 2011.

[51] Jain A. K., Duin R. P. W., and Mao J. Statistical pattern recognition: A review. *IEEE Trans Pattern Anal Mach Intell*, 22:4–37, 2000.

[52] Karssemeijer N. and te Brake G. M. Detection of stellate distortions in mammograms. *IEEE Trans Med Imaging*, 15:611–619, 1996.

[53] Castellano G., Bonilha L., Li L. M., and Cendes F. Texture analysis of medical images. *Clin Radiol*, 59:1061–1069, 2004.

[54] Duda R. O., Hart P. E., and Stork D. G. *Pattern classification*. John Wiley and Sons, New York, 2nd edition, 2001.

[55] Tax D. and Duin R. Using two-class classifiers for multiclass classification. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 124 – 127 vol.2, 2002.

[56] Cortes C. and Vapnik V. Support-vector networks. *Machine Learning*, 20:273–97, 1995.

[57] Freund Y., Shamir E., and Tishby N. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.

[58] Friedman J., Hastie T., and Tibshirani R. Special invited paper. additive logistic regression: A statistical view of boosting. *Ann Stat*, 28:337–374, 2000.

[59] Breiman L. Random forests. *Machine Learning*, 45:5–32, 2001.

[60] Tumer K. and Ghosh J. Error correlation and error reduction in ensemble classifiers. *Connection science*, 8:385–404, 1996.

[61] van Ginneken B., Armato S. G., de Hoop B., van de Vorst S., Duindam T., Niemeijer M., Murphy K., Schilham A. M. R., Retico A., Fantacci M. E., Camarlinghi N., Bagagli F., Gori I., Hara T., Fujita H., Gargano G., Belloti R., Carlo F. D., Megna R., Tangaro S., Bolanos L., Cerello P., Cheran S. C., Torres E. L., and Prokop M. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. *Med Image Anal*, 14:707–722, 2010.

[62] Niemeijer M., Loog M., Abràmoff M. D., Viergever M. A., Prokop M., and van Ginneken B. On combining computer-aided detection systems. *IEEE Trans Med Imaging*, 30:215–223, 2011.

[63] Metz C. E. ROC methodology in radiologic imaging. *Invest Radiol*, 21:720–733, 1986.

[64] Swets J. A. The relative operating characteristic in psychology. *Science*, 182(4116):990–1000, 1973.

[65] Hajian-Tilaki K. O., Hanley J. A., Joseph L., and Collet J.-P. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Med Decis Making*, 17(1):94–102, 1997.

[66] Samulski M., Hupse R., Boetes C., Mus R., den Heeten G., and Karssemeijer N. Using Computer Aided Detection in Mammography as a Decision Support. *Eur Radiol*, 20:2323–2330, 2010.

[67] Efron B. Bootstrap methods: Another look at the jackknife. *Ann Stat*, 7:1–26, 1979.

[68] Samuelson F. and Petrick N. Comparing image detection algorithms using resampling. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, pages 1312 –1315, 2006.

[69] Rutter C. M. Bootstrap estimation of diagnostic accuracy with patient-clustered data. *Acad Radiol*, 7:413–419, 2000.

[70] Dorfman D. D., Berbaum K. S., and Metz C. E. Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. *Invest Radiol*, 27:723–731, 1992.

[71] Becker H. C., Nettleton W. J., Meyers P. H., Sweeney J. W., and Nice Jr. C. M. Digital computer determination of a medical diagnostic index directly from chest X-ray images. *IEEE Trans Biomed Eng*, BME-11:67–72, 1964.

[72] van Ginneken B., ter Haar Romeny B. M., and Viergever M. A. Computer-aided diagnosis in chest radiography: a survey. *IEEE Trans Med Imaging*, 20:1228–1241, 2001.

[73] Katsuragawa S. and Doi K. Computer-aided diagnosis in chest radiography. *Comput Med Imaging Graph*, 31:212–23, 2007.

[74] van Ginneken B., Hogeweg L., and Prokop M. Computer-aided diagnosis in chest radiography: beyond nodules. *Eur J Radiol*, 72:226–230, 2009.

[75] McNitt-Gray M. F., Sayre J. W., Huang H. K., and Razavi M. A pattern classification approach to segmentation of chest radiographs. In *Proceedings of the SPIE*, volume 1898, pages 160–170, 1993.

[76] van Ginneken B., Stegmann M. B., and Loog M. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Med Image Anal*, 10:19–40, 2006.

[77] Seghers D., Loeckx D., Maes F., Vandermeulen D., and Suetens P. Minimal shape and intensity cost path segmentation. *IEEE Trans Med Imaging*, 26:1115–1129, 2007.

[78] De Bruijne M. and Nielsen M. Multi-object segmentation using shape particles. In *Information Processing in Medical Imaging*, pages 59–127. Springer, 2005.

[79] Loog M. and van Ginneken B. Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification. *IEEE Trans Med Imaging*, 25:602–11, 2006.

[80] Simkó G., Orbán G., Máday P., and Horváth G. Elimination of clavicle shadows to help automatic lung nodule detection on chest radiographs. In *4th European Conference of the International Federation for Medical and Biological Engineering*, volume 22 of *IFMBE Proceedings*, pages 488–491. 2009.

[81] McAdams H. P., Samei E., Dobbins, 3rd J., Tourassi G. D., and Ravin C. E. Recent advances in chest radiography. *Radiology*, 241(3):663–683, Dec 2006.

[82] Samei E., Flynn M. J., Peterson E., and Eyler W. R. Subtle lung nodules: influence of local anatomic variations on detection. *Radiology*, 228:76–84, 2003.

[83] Ikeda M., Ishigaki T., and Itoh S. Influence of rib structure on detection of subtle lung nodules. *Eur J Radiol*, 59:49–55, 2006.

[84] Yoshida H. Local contralateral subtraction based on bilateral symmetry of lung for reduction of false positives in computerized detection of pulmonary nodules. *IEEE Trans Biomed Eng*, 51:778–789, 2004.

[85] Freedman M. T., Lo S.-C. B., Seibel J. C., and Bromley C. M. Lung nodules: improved detection with software that suppresses the rib and clavicle on chest radiographs. *Radiology*, 260:265–273, 2011.

[86] Li F., Hara T., Shiraishi J., Engelmann R., MacMahon H., and Doi K. Improved detection of subtle lung nodules by use of chest radiographs with bone suppression imaging: receiver operating characteristic analysis with and without localization. *AJR Am J Roentgenol*, 196: W535–W541, 2011.

[87] Oda S., Awai K., Suzuki K., Yanaga Y., Funama Y., MacMahon H., and Yamashita Y. Performance of radiologists in detection of small pulmonary nodules on chest radiographs: effect of rib suppression with a massive-training artificial neural network. *AJR Am J Roentgenol*, 193:W397–W402, 2009.

[88] Giger M. L., Doi K., MacMahon H., and Metz C. E. Computerized detection of pulmonary nodules in digital chest images: use of morphological filters in reducing false positive

detections. *Med Phys*, 17:861–865, 1990.

[89] Keserci B. and Yoshida H. Computerized detection of pulmonary nodules in chest radiographs based on morphological features and wavelet snake model. *Med Image Anal*, 6: 431–447, 2002.

[90] Suzuki K., Abe H., MacMahon H., and Doi K. Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (MTANN). *IEEE Trans Med Imaging*, 25:406–416, 2006.

[91] Loog M. and van Ginneken B. Bony Structure Suppression in Chest Radiographs. In *Computer Vision Approaches to Medical Image Analysis*, volume 4241 of *Lect Notes Comput Sci*, pages 166–177, 2006.

[92] Chen X., Doi K., Katsuragawa S., and MacMahon H. Automated selection of regions of interest for quantitative analysis of lung textures in digital chest radiographs. *Med Phys*, 20: 975–982, 1993.

[93] Li Q., Katsuragawa S., and Doi K. Improved contralateral subtraction images by use of elastic matching technique. *Med Phys*, 27:1934–1942, 2000.

[94] Li Q., Katsuragawa S., Ishida T., Yoshida H., Tsukuda S., MacMahon H., and Doi K. Contralateral subtraction: A novel technique for detection of asymmetric abnormalities on digital chest radiographs. *Med Phys*, 27:47–55, 2000.

[95] MacMahon H., Liu K. J. M., Montner S. M., and Doi K. The nature and subtlety of abnormal findings in chest radiographs. *Med Phys*, 18:206–210, 1991.

[96] Berlin L. Malpractice and radiologists, update 1986: an 11.5-year perspective. *AJR Am J Roentgenol*, 147(6):1291–1298, 1986.

[97] de Boo D. W., Prokop M., Uffmann M., van Ginneken B., and Schaefer-Prokop C. M. Computer-aided detection (CAD) of lung nodules and small tumours on chest radiographs. *Eur J Radiol*, 72:218–225, 2009.

[98] Sutton R. N. and Hall E. L. Texture measures for automatic classification of pulmonary disease. *IEEE Trans Comput*, 21:667–676, 1972.

[99] Tully R. J., Connors R. W., Harlow C. A., and Lodwick G. S. Towards computer analysis of pulmonary infiltration. *Invest Radiol*, 13:298–305, 1978.

[100] Katsuragawa S., Doi K., and MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography: detection and characterization of interstitial lung disease in digital chest radiographs. *Med Phys*, 15:311–319, 1988.

[101] Kido S., Ikezoe J., Naito H., Tamura S., and Machi S. Fractal analysis of interstitial lung abnormalities in chest radiography. *Radiographics*, 15:1457–1464, 1995.

[102] van Ginneken B., ter Haar Romeny B. M., and Viergever M. A. Automatic segmentation and texture analysis of PA chest radiographs to detect abnormalities related to interstitial disease and tuberculosis. In *Comput Assist Radiol Surg*, pages 685–688, 2002.

[103] Kao E.-F., Kuo Y.-T., Hsu J.-S., Chou M.-C., and Liu G.-C. Zone-based analysis for automa-

ted detection of abnormalities in chest radiographs. *Med Phys*, 38:4241–4251, 2011.

[104] Nakamori N., Doi K., MacMahon H., Sasaki Y., and Montner S. M. Effect of heart-size parameters computed from digital chest radiographs on detection of cardiomegaly: potential usefulness for computer-aided diagnosis. *Invest Radiol*, 26:546–550, 1991.

[105] Coppini G., Miniati M., Monti S., Paterni M., Favilla R., and Ferdeghini E. M. A computer-aided diagnosis approach for emphysema recognition in chest radiography. *Med Eng Phys*, 35:63–73, 2013.

[106] Jaeger S., Karargyris A., Candemir S., Siegelman J., Folio L., Antani S., and Thoma G. Automatic screening for tuberculosis in chest radiographs: a survey. *Quant Imaging Med Surg*, 3(2):89, 2013.

[107] Sarkar S. and Chaudhuri S. Evaluation and progression analysis of pulmonary tuberculosis from digital chest radiographs. *Comput Med Imaging Graph*, 22:145–155, 1998.

[108] Hariharan S., Ray A., and Ghosh M. An algorithm for the enhancement of chest X-ray images of tuberculosis patients. In *Industrial Technology 2000. Proceedings of IEEE International Conference on*, volume 2, pages 107–112, 2000.

[109] Song Y.-L. and Yang Y. Localization algorithm and implementation for focal of pulmonary tuberculosis chest image. In *Machine Vision and Human-Machine Interface (MVHI), 2010 International Conference on*, pages 361–364. IEEE, 2010.

[110] Koeslag A. and de Jager G. Computer aided diagnosis of miliary tuberculosis. *Proceedings of the Pattern Recognition Association of South Africa*, 2001.

[111] van Ginneken B., Katsuragawa S., ter Haar Romeny B. M., Doi K., and Viergever M. A. Automatic detection of abnormalities in chest radiographs using local texture analysis. *IEEE Trans Med Imaging*, 21:139–149, 2002.

[112] Arzhaeva Y., Hogeweg L., de Jong P. A., Viergever M. A., and van Ginneken B. Global and Local Multi-valued Dissimilarity-Based Classification: Application to Computer-Aided Detection of Tuberculosis. In *Med Image Comput Comput Assist Interv*, Lect Notes Comput Sci, pages 724–731, 2009.

[113] Lieberman R., Kwong H., Liu B., and Huang H. Computer-assisted detection (CAD) methodology for early detection of response to pharmaceutical therapy in tuberculosis patients. *Proc Soc Photo Opt Instrum Eng*, 7260:726030, 2009.

[114] Shen R., Cheng I., and Basu A. A hybrid knowledge-guided detection technique for screening of infectious pulmonary tuberculosis from chest radiographs. *IEEE Trans Biomed Eng*, 57(11):2646–2656, 2010.

[115] Tan J. H., Acharya U. R., Tan C., Abraham K. T., and Lim C. M. Computer-assisted diagnosis of tuberculosis: A first order statistical approach to chest radiograph. *J Med Syst*, 36(5):2751–2759, 2012.

[116] Jaeger S., Karargyris A., Antani S., and Thoma G. Detecting tuberculosis in radiographs using combined lung masks. In *Engineering in Medicine and Biology Society (EMBC), 2012*

Annual International Conference of the IEEE, pages 4978–4981, 2012.

[117] Niemeijer M., Abràmoff M. D., and van Ginneken B. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. *Med Image Anal*, 10:888–898, 2006.

[118] Yoon S., Feng J., and Jain A. K. Altered fingerprints: analysis and detection. *IEEE Trans Pattern Anal Mach Intell*, 34:451–464, 2012.

[119] Keller B. M., Reeves A. P., Cham M. D., Henschke C. I., and Yankelevitz D. F. Semi-automated location identification of catheters in digital chest radiographs. In *SPIE Medical Imaging*, volume 6514, pages 65141O–65141O, March 2007.

[120] Rudin L., Osher S., and Fatemi E. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259 – 268, 1992.

[121] Kuan D., Sawchuk A., Strand T., and Chavel P. Adaptive restoration of images with speckle. *IEEE Trans Acoust Speech Signal Process*, 35:373 – 383, 1987.

[122] Lee T., Ng V., Gallagher R., Coldman A., and McLean D. Dullrazor : A software approach to hair removal from images. *Comput Biol Med*, 27:533 – 543, 1997.

[123] Efros A. A. and Leung T. K. Texture synthesis by non-parametric sampling. In *Proc. Seventh IEEE Int Computer Vision Conf.*, volume 2, pages 1033–1038, 1999.

[124] Bertalmio M., Sapiro G., Caselles V., and Ballester C. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '00, pages 417–424, 2000.

[125] Emile-Male G. *The restorer's handbook of easel painting*. Van Nostrand Reinhold, 1st edition, 1976.

[126] Bertalmio M., Vese L., Sapiro G., and Osher S. Simultaneous structure and texture image inpainting. *IEEE Trans Image Process*, 12:882 – 889, 2003.

[127] Criminisi A., Perez P., and Toyama K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans Image Process*, 13:1200 –1212, 2004.

[128] Zhou H., Chen M., Gass R., Rehg J. M., Ferris L., Ho J., and Drogowski L. Feature-preserving artifact removal from dermoscopy images. In *SPIE Conference Series*, volume 6914, April 2008.

[129] Wighton P., Lee T. K., and Atkins M. S. Dermascopic hair disocclusion using inpainting. In *Medical Imaging*, Proceedings of the SPIE, pages 691427 – 691427–8, 2008.

[130] Bornemann F. and März T. Fast image inpainting based on coherence transport. *J Math Imaging Vis*, 28:259–278, 2007.

[131] Abbas Q., Garcia I., Celebi M., and Ahmad W. A feature-preserving hair removal algorithm for dermoscopy images. *Skin Research and Technology*, pages 1–10, 2011.

[132] Schilham A. M. R., van Ginneken B., and Loog M. A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database.

Med Image Anal, 10:247–258, 2006.

[133] Florack L. M. J., ter Haar Romeny B. M., Viergever M. A., and Koenderink J. J. The Gaussian scale-space paradigm and the multiscale local jet. *Int J Comput Vis*, 18:61–75, 1996.

[134] Deriche R. Fast algorithms for low-level vision. *IEEE Trans Pattern Anal Mach Intell*, 12:78 –87, 1990.

[135] Frangi A. F., Niessen W. J., Vincken K. L., and Viergever M. A. Multiscale vessel enhancement filtering. In *Med Image Comput Comput Assist Interv*, volume 1496 of *Lect Notes Comput Sci*, pages 130–137, 1998.

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

[137] Torralba A., Murphy K., and Freeman W. Sharing visual features for multiclass and multiview object detection. *IEEE Trans Pattern Anal Mach Intell*, 29:854–869, 2007.

[138] Lienhart R., Kuranov A., and Pisarevsky V. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Pattern Recognition*, volume 2781 of *Lect Notes Comput Sci*, pages 297–304. 2003.

[139] Jain A. K. and Li S. Z. *Handbook of Face Recognition*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 038740595X.

[140] Chang C.-C. and Lin C.-J. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol*, 2:27:1–27:27, May 2011.

[141] Wei L. and Levoy M. Fast texture synthesis using tree-structured vector quantization. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '00, pages 479–488, 2000.

[142] Hsu C.-W., Chang C.-C., and Lin C.-J. A practical guide to support vector classification, 2010. URL `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

[143] Pudil P., Novovicova J., and Kittler J. Floating search methods in feature selection. *Pattern Recognit Lett*, 15:1119–1125, 1994.

[144] Bunch P., Hamilton J., Sanderson G., and Simmons A. A free response approach to the measurement and characterization of radiographic-observer performance. *J Appl Photogr Eng*, 4:166–172, 1978.

[145] Zwiggelaar R., Taylor C., and Rubin C. Detection of the central mass of spiculated lesions – signature normalisation and model data aspects. In *Information Processing in Medical Imaging*, volume 1613 of *Lect Notes Comput Sci*, pages 406–411. 1999.

[146] Oliver A., Freixenet J., Martí J., Pérez E., Pont J., Denton E., and Zwiggelaar R. A review of automatic mass detection and segmentation in mammographic images. *Med Image Anal*, 14:87–110, 2010.

[147] Arzhaeva Y., Prokop M., Tax D. M. J., de Jong P. A., Schaefer-Prokop C. M., and van Ginneken B. Computer-aided detection of interstitial abnormalities in chest radiographs using a reference standard based on computed tomography. *Med Phys*, 34:4798–4809, 2007.

[148] Hogeweg L., Mol C., de Jong P. A., Dawson R., Ayles H., and van Ginneken B. Fusion of local and global detection systems to detect tuberculosis in chest radiographs. In *Med Image Comput Comput Assist Interv*, volume 6363 of *Lect Notes Comput Sci*, pages 650–657, 2010.

[149] Loog M. and van Ginneken B. 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, 2004.

[150] McGill R., Tukey J. W., and Larsen W. A. Variations of box plots. *Am Stat*, 32:12–16, 1978.

[151] Quekel L. G., Kessels A. G., Goei R., and Engelshoven J. M. v. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest*, 115:720–724, 1999.

[152] Yu T., Luo J., and Ahuja N. Shape regularized active contour using iterative global search and local optimization. In *Computer Vision and Pattern Recognition*, volume 2, pages 655–662, 2005.

[153] Cootes T. F., Taylor C. J., Cooper D., and Graham J. Active shape models – their training and application. *Comput Vis Image Underst*, 61:38–59, 1995.

[154] van Rikxoort E. M., van Ginneken B., Klik M., and Prokop M. Supervised enhancement filters: application to fissure detection in chest CT scans. *IEEE Trans Med Imaging*, 27:1–10, 2008.

[155] van Ginneken B. Supervised probabilistic segmentation of pulmonary nodules in CT scans. In *Med Image Comput Comput Assist Interv*, volume 4191 of *Lect Notes Comput Sci*, pages 912–919, 2006.

[156] Cootes T. F. and Taylor C. J. Statistical models of appearance for computer vision. Technical report, 2001.

[157] Vittitoe N. F., Vargas-Voracek R., and Floyd Jr. C. E. Markov random field modeling in posteroanterior chest radiograph segmentation. *Med Phys*, 26:1670–1677, 1999.

[158] Bellman R. E. *Applied Dynamic Programming*. Princeton University Press, 1962.

[159] Montanari U. On the optimal detection of curves in noisy pictures. *Commun ACM*, 14: 335–345, 1971.

[160] Blum H. A transformation for extracting new descriptors of shape. *Models for the Perception of Speech and Visual Form*, pages 362–380, 1967.

[161] Gerig G., Jomier M., and Chakos M. Valmet: a new validation tool for assessing and improving 3D object segmentation. In *Med Image Comput Comput Assist Interv*, Lect Notes Comput Sci, pages 516–523, 2001.

[162] Timp S. and Karssemeijer N. A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography. *Med Phys*, 31:958–971, 2004.

[163] Cui J., Sahiner B., Chan H., Nees A., Paramagul C., Hadjiiski L., Zhou C., and Shi J. A new automated method for the segmentation and characterization of breast masses on ultrasound images. *Med Phys*, 36:1553–1565, 2009.

[164] van Rikxoort E. M., de Hoop B., Viergever M. A., Prokop M., and van Ginneken B. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Med Phys*, 36:2934–2947, 2009.

[165] Warfield S. K., Zou K. H., and Wells W. M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*, 23:903 –921, 2004.

[166] Artaechevarria X., Muñoz-Barrutia A., and de Solórzano C. O. Combination strategies in multi-atlas image segmentation: Application to brain MR data. *IEEE Trans Med Imaging*, 28 (8):1266–1277, 2009.

[167] Kuncheva L. I. A theoretical study on six classifier fusion strategies. *IEEE Trans Pattern Anal Mach Intell*, 24:281–286, 2002.

[168] Brejl M. and Sonka M. Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples. *IEEE Trans Med Imaging*, 19:973–985, 2000.

[169] Deriche R. Recursively implementating the Gaussian and its derivatives. Rapport de recherche, 1993. URL `http://hal.inria.fr/inria-00074778/en/`.

[170] Viola P. and Jones M. Rapid object detection using a boosted cascade of simple features. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages I–511 – I–518, 2001.

[171] Dam E. B. and Loog M. Efficient segmentation by sparse pixel classification. *IEEE Trans Med Imaging*, 27:1525 –1534, 2008.

[172] World Health Organization. WHO report 2008: Global tuberculosis control, Surveillance, Planning, Financing, 2008.

[173] Shah P. K., Austin J. H. M., White C. S., Patel P., Haramati L. B., Pearson G. D. N., Shiau M. C., and Berkmen Y. M. Missed non-small cell lung cancer: Radiographic findings of potentially resectable lesions evident only in retrospect. *Radiology*, 226:235–241, 2003.

[174] Chen S. and Suzuki K. Computerized detection of lung nodules by means of "virtual dual-energy"radiography. *IEEE Trans Biomed Eng*, 60:369–378, 2013.

[175] Suzuki K., Armato III S. G., Engelmann R., Caligiuri P., and MacMahon H. Temporal subtraction of 'virtual dual-energy' chest radiographs for improved conspicuity of growing cancers and other pathologic changes. In *Medical Imaging*, Proceedings of the SPIE, pages 79630F–79630F–6, 2011.

[176] Wang X. and Poor H. Blind multiuser detection: A subspace approach. *IEEE Trans Inf Theory*, 44:677–690, 1998.

[177] Asano F., Ikeda S., Ogawa M., Asoh H., and Kitawaki N. Combined approach of array processing and independent component analysis for blind separation of acoustic signals. *IEEE Trans Speech Audio Process*, 11:204–215, 2003.

[178] Vorobyov S. and Cichocki A. Blind noise reduction for multisensory signals using ICA and

subspace filtering, with application to EEG analysis. *Biol Cybern*, 86:293–303, 2002.

[179] Cichocki A., Shishkin S., Musha T., Leonowicz Z., Asada T., and Kurachi T. EEG filtering based on blind source separation (BSS) for early detection of alzheimer's disease. *Clin Neurophysiol*, 116:729–737, 2005.

[180] Bookstein F. L. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans Pattern Anal Mach Intell*, 11:567–585, 1989.

[181] Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[182] Comon P. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[183] Lee D. D. and Seung H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[184] Automatic nodule detection 2009 (ANODE09). http://anode09.isi.uu.nl/, 2009.

[185] van Rikxoort E. M. and van Ginneken B. Automatic segmentation of the lungs and lobes from thoracic CT scans. In *The Fourth International Workshop on Pulmonary Image Analysis*, pages 261–268, 2011.

[186] Staal J. J. *Segmentation of elongated structures in medical images*. PhD thesis, Utrecht University, The Netherlands, 2004.

[187] Hogeweg L., Sánchez C. I., de Jong P. A., Maduskar P., and van Ginneken B. Clavicle segmentation in chest radiographs. *Med Image Anal*, 16:1490 – 1502, 2012.

[188] Hogeweg L., Mol C., de Jong P. A., and van Ginneken B. Rib suppression in chest radiographs to improve classification of textural abnormalities. In *Medical Imaging*, volume 7624 of *Proceedings of the SPIE*, pages 76240Y1–76240Y6, 2010.

[189] Snoeren P. R., Litjens G. J. S., van Ginneken B., and Karssemeijer N. Training a computer aided detection system with simulated lung nodules in chest radiographs. In *The Third International Workshop on Pulmonary Image Analysis*, pages 139–149, 2010.

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

[191] Kuhnigk J. M., Dicken V., Bornemann L., Bakai A., Wormanns D., Krass S., and Peitgen H. O. Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. *IEEE Trans Med Imaging*, 25:417–434, 2006.

[192] de Souza P. Automatic rib detection in chest radiographs. *Comput Vis Graph Image Process*, 23:129–161, 1983.

[193] Powell G. F., Doi K., and Katsuragawa S. Localization of inter-rib spaces for lung texture analysis and computer-aided diagnosis in digital chest images. *Med Phys*, 15:581–587, 1988.

[194] Sanada S., Doi K., and MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography: automated delineation of posterior ribs in chest images. *Med Phys*, 18:964–971, 1991.

[195] Yue Z., Goshtasby A., and Ackerman L. V. Automatic detection of rib borders in chest radiographs. *IEEE Trans Med Imaging*, 14:525–536, 1995.

[196] Chen S. and Suzuki K. Bone suppression in chest radiographs by means of anatomically specific multiple massive-training ANNs. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 17–20, 2012.

[197] Ramakrishna B., Brown M., Goldin J., Cagnon C., and Enzmann D. An improved automatic computer aided tube detection and labeling system on chest radiographs. In *Medical Imaging*, volume 8315 of *Proceedings of the SPIE*, pages 83150R–1, 2012.

[198] Hogeweg L., Sánchez C. I., Melendez J., Maduskar P., Story A., Hayward A., and van Ginneken B. Foreign object detection and removal to improve automated analysis of chest radiographs. *Med Phys*, 40(7):071901, 2013.

[199] Kuhlman J. E., Collins J., Brooks G. N., Yandow D. R., and Broderick L. S. Dual-energy subtraction chest radiography: what to look for beyond calcified nodules. *Radiographics*, 26:79–92, 2006.

[200] Laguna P., Moody G., Garcia J., Goldberger A., and Mark R. Analysis of the ST-T complex of the electrocardiogram using the Karhunen-Loeve transform: adaptive monitoring and alternans detection. *Med Biol Eng Comput*, 37:175–189, 1999.

[201] Sadasivan P. K. and Dutt D. N. SVD based technique for noise reduction in electroencephalographic signals. *Signal Processing*, 55:179–189, 1996.

[202] Beckmann C. F. and Smith S. M. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans Med Imaging*, 23:137–152, 2004.

[203] Huber P. J. *Robust Statistics*. John Wiley & Sons, New York, 1981.

[204] Saber E. and Tekalp A. M. Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. *Pattern Recognition Letters*, 19(8):669–680, 1998.

[205] Li W. H. and Kleeman L. Real time object tracking using reflectional symmetry and motion. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 2798–2803, 2006.

[206] Javed O. and Shah M. Tracking and object classification for automated surveillance. In *Computer Vision–ECCV 2002*, pages 343–357. Springer, 2006.

[207] Hays J., Leordeanu M., Efros A. A., and Liu Y. Discovering texture regularity as a higher-order correspondence problem. In *Computer Vision–ECCV 2006*, pages 522–535. Springer, 2006.

[208] Liu Y., Hel-Or H., and Kaplan C. S. *Computational symmetry in computer vision and computer graphics*. Now publishers Inc, 2010.

[209] Marola G. On the detection of the axes of symmetry of symmetric and almost symmetric planar images. *IEEE Trans Pattern Anal Mach Intell*, 11:104–108, 1989.

[210] Zabrodsky H., Peleg S., and Avnir D. Symmetry as a continuous feature. *IEEE Trans Pattern Anal Mach Intell*, 17:1154–1166, 1995.

[211] Giblin P. J. and Kimia B. B. On the intrinsic reconstruction of shape from its symmetries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(7):895–911, July 2003.

[212] Scognamillo R., Rhodes G., Morrone C., and Burr D. A feature–based model of symmetry detection. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1525): 1727–1733, 2003.

[213] Loy G. and Eklundh J.-O. Detecting symmetry and symmetric constellations of features. In *Computer Vision–ECCV 2006*, pages 508–521. Springer, 2006.

[214] Grammer K., Thornhill R., et al. Human (homo sapiens) facial attractiveness and sexual selection: the role of symmetry and averageness. *Journal of comparative psychology*, 108(3): 233–242, 1994.

[215] Graham J. H., Raz S., Hel-Or H., and Nevo E. Fluctuating asymmetry: methods, theory, and applications. *Symmetry*, 2(2):466–540, 2010.

[216] Keinan S. and Avnir D. Quantitative symmetry in structure-activity correlations: The near C2 symmetry of inhibitor/HIV protease complexes. *Journal of the American Chemical Society*, 122(18):4378–4384, 2000.

[217] Liu Y., Collins R. T., and Rothfus W. E. Robust midsagittal plane extraction from normal and pathological 3-D neuroradiology images. *IEEE Trans Med Imaging*, 20:175–192, 2001.

[218] Sun Y., Bhanu B., and Bhanu S. Automatic symmetry-integrated brain injury detection in MRI sequences. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 79–86, 2009.

[219] Sun Y. and Bhanu B. Reflection symmetry-integrated image segmentation. *IEEE Trans Pattern Anal Mach Intell*, 34(9):1827–1841, Sep 2012.

[220] Kowner R. and Thornhill R. The imperfect organism: On the concept of fluctuating asymmetry and its significance in human, non-human animals, and plants. *Symmetry: Culture and Science*, 10(3-4):227–244, 1999.

[221] Good C. D., Johnsrude I., Ashburner J., Henson R. N., Friston K. J., and Frackowiak R. S. Cerebral asymmetry and the effects of sex and handedness on brain structure: a voxel-based morphometric analysis of 465 normal adult human brains. *Neuroimage*, 14(3):685–700, 2001.

[222] Lowe D. G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157, 1999.

[223] Sivic J. and Zisserman A. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages

1470–1477. IEEE, 2003.

[224] Avni U., Greenspan H., Konen E., Sharon M., and Goldberger J. X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. *IEEE Trans Med Imaging*, 30(3):733–746, 2011.

[225] Ojala T., Pietikainen M., and Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell*, 24: 971–987, 2002.

[226] Liu T., Moore A. W., Gray A., and Yang K. An investigation of practical approximate nearest neighbor algorithms. *Advances in neural information processing systems*, 17:825–832, 2004.

[227] Kundel H. L. and Nodine C. F. Interpreting chest radiographs without visual search. *Radiology*, 116:527–532, 1975.

[228] Koontz N. A. and Gunderman R. B. Gestalt theory: implications for radiology education. *AJR Am J Roentgenol*, 190:1156–1160, 2008.

[229] Mikolajczyk K. and Schmid C. A performance evaluation of local descriptors. *IEEE Trans Pattern Anal Mach Intell*, 27:1615–1630, 2005.

[230] Glickman M. S. and Jacobs Jr W. R. Microbial pathogenesis review of mycobacterium tuberculosis: Dawn of a discipline. *Cell*, 104:477–485, 2001.

[231] Ishida T., Katsuragawa S., Kobeyashi T., MacMahon H., and Doi K. Computerized analysis of interstitial disease in chest radiographs: improvement of geometric-pattern feature analysis. *Med Phys*, 24:915–924, 1997.

[232] van Ginneken B., Frangi A. F., Staal J. J., ter Haar Romeny B. M., and Viergever M. A. Active shape model segmentation with optimal features. *IEEE Trans Med Imaging*, 21:924–933, 2002.

[233] Arzhaeva Y., Tax D. M. J., and van Ginneken B. Dissimilarity-based classification in the absence of local ground truth: application to the diagnostic interpretation of chest radiographs. *Pattern Recognit*, 42:1768–1776, 2009.

[234] Kuncheva L. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[235] Kittler J., Hatef M., Duin R. P. W., and Matas J. On combining classifiers. *IEEE Trans Pattern Anal Mach Intell*, 20:226–239, 1998.

[236] Tax D. *One-class classification*. PhD thesis, Delft University of Technology, 2001.

[237] Duin R. The combining classifier: To train or not to train? In *International Conference on Pattern Recognition*, volume 16, pages 765–770, 2002.

[238] Evaluation of multiple novel and emerging technologies for TB diagnosis, in smear-negative and HIV-infected persons, in high burden countries (the TB-NEAT study). URL `http://www.lunginstitute.co.za/tbneat/`.

[239] CRRS guidelines. URL `http://www.lunginstitute.co.za/downloads/CRRSGuidelines.pdf`.

[240] Ho T. and Basu M. Complexity measures of supervised classification problems. *IEEE Trans Pattern Anal Mach Intell*, 24:289–300, 2002.

[241] Woodring J. H. and Mac Vandiviere H. Pulmonary disease caused by nontuberculous mycobacteria. *J Thorac Imaging*, 5:64–76, 1990.

[242] Levy H., Feldman C., Sacho H., Van Der Meulen H., Kallenbach J., and Koornhof H. A reevaluation of sputum microscopy and culture in the diagnosis of pulmonary tuberculosis. *CHEST Journal*, 95:1193–1197, 1989.

[243] Armato S. G., Giger M. L., and MacMahon H. Computerized delineation and analysis of costophrenic angles in digital chest radiographs. *Acad Radiol*, 5:329–335, 1998.

[244] Theron G., Pooran A., Peter J., van Zyl-Smit R., Mishra H. K., Meldau R., Calligaro G., Allwood B., Sharma S. K., Dawson R., and Dheda K. Do adjunct TB tests, when combined with Xpert MTB/RIF, improve accuracy and the cost of diagnosis in a resource-poor setting? *Eur Respir J*, 40(1):161–168, 2011.

[245] Graham S., Das G. K., Hidvegi R., Hanson R., Kosiuk J., Al Z., Menzies D., et al. Chest radiograph abnormalities associated with tuberculosis: reproducibility and yield of active cases. *Int J Tuberc Lung Dis*, 6(2):137, 2002.

[246] Abubakar I., Story A., Lipman M., Bothamley G., van Hest R., Andrews N., Watson J. M., and Hayward A. Diagnostic accuracy of digital chest radiography for pulmonary tuberculosis in a uk urban population. *Eur Respir J*, 35:689–692, 2010.

[247] Gilbert F. J., Astley S. M., Gillan M. G. C., Agbaje O. F., Wallis M. G., James J., Boggis C. R. M., Duffy S. W., and CADET II Group. Single reading with computer-aided detection for screening mammography. *N Engl J Med*, 359:1675–1684, 2008.

[248] Murphy K., van Ginneken B., Schilham A. M. R., de Hoop B. J., Gietema H. A., and Prokop M. A large scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Med Image Anal*, 13:757–770, 2009.

[249] Regge D., Della Monica P., Galatola G., Laudi C., Zambon A., Correale L., Asnaghi R., Barbaro B., Borghi C., Campanella D., et al. Efficacy of computer-aided detection as a second reader for 6–9-mm lesions at CT colonography: Multicenter prospective trial. *Radiology*, 266 (1):168–176, 2013.

[250] Abràmoff M. D., Niemeijer M., Suttorp-Schulten M. S. A., Viergever M. A., Russell S. R., and van Ginneken B. Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes. *Diabetes Care*, 31:193–198, 2008.

[251] Chan H. P., Doi K., Vyborny C. J., Schmidt R. A., Metz C. E., Lam K. L., Ogura T., Wu Y. Z., and MacMahon H. Improvement in radiologists' detection of clustered microcalcifications on mammograms. the potential of computer-aided diagnosis. *Invest Radiol*, 25:1102–1110, 1990.

[252] Uppaluri R., Hoffman E. A., Sonka M., Hartley P. G., Hunninghake G. W., and McLennan

G. Computer recognition of regional lung disease patterns. *Am J Respir Crit Care Med*, 160: 648–654, 1999.

[253] Doi K. Computer-aided diagnosis and its potential impact on diagnostic radiology. In *Computer-aided diagnosis in medical imaging*. Elsevier, 1999.

[254] Lieberman D. A., Harford W. V., Ahnen D. J., Provenzale D., Sontag S. J., Schnell T. G., Chejfec G., Campbell D. R., Durbin T. E., Bond J. H., et al. One-time screening for colorectal cancer with combined fecal occult-blood testing and examination of the distal colon. *N Engl J Med*, 345(8):555–560, 2001.

[255] Catalona W. J., Smith D. S., Ratliff T. L., Dodds K. M., Coplen D. E., Yuan J. J., Petros J. A., and Andriole G. L. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *N Engl J Med*, 324:1156–1161, 1991.

[256] Health Protection Agency. Tuberculosis in the UK: 2012 report. URL `http://www.hpa.org.uk/Publications/InfectiousDiseases/Tuberculosis/1206TBintheUK2012report`.

[257] UCL Research Ethics Committee. Research ethics at UCL. URL `http://ethics.grad.ucl.ac.uk/forms/leaflet.pdf`.

[258] Jit M., Stagg H. R., Aldridge R. W., White P. J., and Abubakar I. Dedicated outreach service for hard to reach patients with tuberculosis in london: observational study and economic evaluation. *British Medical Journal*, 343, 2011.

[259] World Health Organization. Tuberculosis prevalence surveys: a handbook. World Health Organization, 2011. URL `http://whqlibdoc.who.int/publications/2011/9789241548168_eng.pdf`.

[260] Balassy C., Prokop M., Weber M., Sailer J., Herold C. J., and Schaefer-Prokop C. Flat-panel display (LCD) versus high-resolution gray-scale display (CRT) for chest radiography: an observer preference study. *AJR Am J Roentgenol*, 184:752–756, 2005.

[261] Zellweger J. P., Heinzer R., Touray M., Vidondo B., and Altpeter E. Intra-observer and overall agreement in the radiological assessment of tuberculosis. *Int J Tuberc Lung Dis*, 10: 1123–1126, 2006.

[262] World Health Organization. WHO report 2009: Global tuberculosis control, epidemiology, strategy, financing, 2009.

[263] Duin R. and Tax D. Experiments with classifier combining rules. *Multiple Classifier Systems*, pages 16–29, 2000.

[264] Ralph A. P., Ardian M., Wiguna A., Maguire G. P., Becker N. G., Drogumuller G., Wilks M. J., Waramori G., Tjitra E., Sandjaja, Kenagalem E., Pontororing G. J., Anstey N. M., and Kelly P. M. A simple, valid, numerical score for grading chest x-ray severity in adult smear-positive pulmonary tuberculosis. *Thorax*, 65:863–869, 2010.

[265] Pai N. P., Vadnais C., Denkinger C., Engel N., and Pai M. Point-of-care testing for infectious diseases: diversity, complexity, and barriers in low-and middle-income countries. *PLoS Med*, 9(9):e1001306, 2012.

[266] Walker D. Economic analysis of tuberculosis diagnostic tests in disease control: how can it be modelled and what additional information is needed? *Int J Tuberc Lung Dis*, 5(12): 1099–1108, 2001.

[267] Corbett E. L., Bandason T., Duong T., Dauya E., Makamure B., Churchyard G. J., Williams B. G., Munyati S. S., Butterworth A. E., Mason P. R., et al. Comparison of two active case-finding strategies for community-based diagnosis of symptomatic smear-positive tuberculosis and control of infectious tuberculosis in Harare, Zimbabwe (DETECTB): a cluster-randomised trial. *Lancet*, 376(9748):1244, 2010.

[268] Maduskar P., Hogeweg L., Philipsen R., and van Ginneken B. Automated localization of costophrenic recesses and costophrenic angle measurement on frontal chest radiographs. In *Medical Imaging*, volume 8670 of *Proceedings of the SPIE*, page 867038, 2013.

[269] Okumura E., Kawashita I., and Ishida T. Computerized analysis of pneumoconiosis in digital chest radiography: Effect of artificial neural network trained with power spectra. *J Digit Imaging*, 24(6):1126–1132, 2010.

[270] Yu P., Xu H., Zhu Y., Yang C., Sun X., and Zhao J. An automatic computer-aided detection scheme for pneumoconiosis on digital chest radiographs. *J Digit Imaging*, 24(3):382–393, 2011.

[271] Settles B. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, San Rafael, 2012.

[272] Cohn D. A., Ghahramani Z., and Jordan M. I. Active learning with statistical models. *J Artif Intell Res*, 4:129–145, 1996.

[273] Acuna E. and Rodriguez C. The treatment of missing values and its effect on classifier accuracy. In *Classification, Clustering, and Data Mining Applications*, pages 639–647. Springer, 2004.

[274] Pan S. J. and Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010.

[275] Kallenberg M. and Karssemeijer N. Computer-aided detection of masses in full-field digital mammography using screen-film mammograms for training. *Phys Med Biol*, 53:6879–6891, 2008.

[276] Philipsen R., Maduskar P., Hogeweg L., and van Ginneken B. Normalization of chest radiographs. In *Medical Imaging*, volume 8670 of *Proceedings of the SPIE*, page 86700G, 2013.

[277] Wilson D., Nachega J., Morroni C., Chaisson R., and Maartens G. Diagnosing smear-negative tuberculosis using case definitions and treatment response in HIV-infected adults. *Int J Tuberc Lung Dis*, 10(1):31–38, 2006.

[278] Pinto L. M., Dheda K., Theron G., Allwood B., Calligaro G., van Zyl-Smit R., Peter J., Schwartzman K., Menzies D., Bateman E., Pai M., and Dawson R. Development of a simple reliable radiographic scoring system to aid the diagnosis of pulmonary tuberculosis. *PLoS One*, 8:e54235, 2013.

[279] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[280] Le Q., Ranzato M., Monga R., Devin M., Chen K., Corrado G., Dean J., and Ng A. Building high-level features using large scale unsupervised learning. In *International Conference in Machine Learning*, 2012.

[281] Dreyfus H. L. *What computers still can't do - a critique of artificial reason*. MIT Press, Cambridge, 1992. ISBN 978-0-262-04134-8.

Acknowledgements

First a little bit of history. After my master's degrees it seemed logical to pursue a PhD that combined image processing and medicine. I already had an interest in the topic of *machine learning* and when I found out about Bram van Ginneken's Computer Aided Diagnosis Group, this was the perfect choice. I remember vividly the first appointment with Bram to discuss project possibilities where we discussed all kinds of topics, which, in hindsight, showed a slight naivety on my side. Bram did not have any open projects at that moment, but advised me to look for options in Nijmegen with prof. Nico Karssemeijer. Luckily a few months later Bram contacted me again for a project on chest X-ray and tuberculosis in Utrecht.

Bram, I would like to thank you for being an excellent supervisor and promotor and for giving me the chance to work in your group. I've rarely encountered anyone who is more motivated and as deeply involved in the subjects he is working on. I also really appreciate your knowledge of both the smallest technical details and the big picture. You have provided many of the ideas that can be found in this thesis, originating from countless interesting discussions. Clarisa Sánchez, many thanks for being my co-promotor and helping me in the last two years with the study design and the writing of the papers. Your preference for preciseness has significantly improved this thesis. Prof. Mathias Prokop, I would like to thank you for your seemingly endless knowledge on the topic of (lung) radiology and for welcoming me in the radiology department in Utrecht. Prof. Max Viergever, thank you for creating such a great group on image processing and welcoming me there.

Prof. dr. ir. Nico Verdonschot, prof. dr. ir. Bart ter Haar Romeny, and dr. Martin Boeree, I wish to thank you for participation in the reading committee.

A word of thanks should go to my co-authors as well. Pim de Jong, thanks for being a great tutor in chest X-ray analysis. I wish we could have worked more together more often, but unfortunately the rapidly expanding number of topics in the CAD4TB project prevented this. Rodney Dawson, Helen Ayles, Grant Theron, and prof. Keertan Dheda, thank you for collaborating on several of the papers in this thesis and for allowing me to join you in Africa and get a direct feeling of the huge problem that TB still is in the world. Al Story, Rob Aldridge, Andrew Hayward, and prof. Ibrahim Abubakar, thanks for welcoming me to London and for your support in writing the final chapter of this thesis.

This project would have been largely impossible without the support of Oldelft, who provided the X-ray cameras that were used to acquire many of the images

used in this thesis. Frank Vijn, I enjoyed working with you, your knowledge about practical TB (research) in Africa has been very educational and motivating. I really enjoyed the visit to Zambia where you showed us the difficulties in acquiring the chest radiographs that were the subject of several papers in this thesis. Guido Geerts, thanks for keeping faith in this project, despite thinks seemly going a bit slow sometimes. Frank van Doren and Wessel Eijkman, I profited a lot from your experience, enthusiasm and many questions.

The majority of my PhD was spent at the Diagnostic Image Analysis Group. I've seen the group grow since its inception and I think it's a great place to do research. As a result of the large amount of knowledge which is available, but also because of the many friendly people who make the group what it is. I fondly remember playing soccer, table tennis, the drinks at Café Samson and the infamous DIAG weekends. Michiel, you were one of my first roommates when I moved to Nijmegen. When I saw your "cruesli wall" and the organization of your desk, I knew we had something in common. Thanks, for endless discussions on arbitrary subjects, grumblings about doing a PhD and cooking birds. Also thanks for inviting me to the Wednesday's pub quiz. Geert, I already knew you as an enthusiastic student when you did your internship in the CAD4TB project back in Utrecht. I think it's great you joined DIAG. Eva, thanks for your inexhaustible cheerfulness. Although we have not worked together much in research, I enjoy being part of the Chest CT project as a scientific programmer. Rick, you only joined the CAD4TB team in the last year, but in that time I have enjoyed working with you a lot. I feel I leave the project in safe hands with you and Pragnya.

Special thanks to my paranymphs, Sjoerd en Pragnya. Pragnya, I think it is great to have an international team member on board in the CAD4TB project. I enjoyed all the delicacies you brought into the office. We shared many insights, but also frustrations during the three years as roommates. Good luck with finishing your thesis. Sjoerd, I will miss your ability to see things in perspective, your feeling for absurdness, our discussions about almost anything and your interest in philosophy.

The first year of my PhD I was part of the (CAD) group in Utrecht. I would like to thank everybody there for a great time. Christian, you're one of the most skillful programmers I have ever met and I owe my C++ skills largely to you. I hope we could share our tastes for whisky, Skyrim, and Buurman & Buurman more often. Ewoud "Google" Smit, I had a great time with you as a roommate. I hope you finish your PhD soon. Thessa, Steven, Thomas, Ivana, Adriënne, and

Sascha, I enjoyed your company in the group and during drinks and other activities.

Next to all the people directly involved in my thesis I would like to thank my friends for their support during the four years. Wouter and Ragnhild, thanks for your great company and welcoming us in your beautiful country. Remco, Jenne, Jeroen, Jan, Erik "the Viking", Teun and Remko, thanks for the yearly Christmas "LAN"; playing computer games and drinking beer was much needed in order to recharge for writing a thesis. Jenne, and Remco, your artistic view of the world kept things in perspective for me. This holds especially true for the philosophical (and other) discussions with Erik; these other-worldly digressions were a welcome distraction at times. Jan en Janwillem, your professional attitude and motivation in your own fields of expertise have been an example to me. Tobi, Anneleen, Lubine, Tim, and other pubquizzers, I thoroughly enjoyed your weekly company (and Tobi's knowledge on metal music).

Arend, Marjan, Henk, Saskia, and Vera-Lotte thanks for welcoming me in your family. Mom, dad, Cleo, Judith and Yara, thanks for all the support and interest for what I have been doing.

Joanna, you have had to endure the most in those four years. Thanks for keeping me motivated and being so sweet in both happy and difficult times. I'm very glad to be with you. Luna, you have been an important motivation to finish this thesis, even before you surfaced the earth.

Special thanks to Trent, Ludwig, Friedrich, John(n), and Simeon.

Curriculum Vitae

Curriculum Vitae

Laurens Hogeweg was born in Hoorn, the Netherlands, on 17 Augustus 1982. After high school at the Openbare Scholengemeenschap (Hoorn) in 2000, he went on to study medicine at the Rijksuniversiteit Groningen (RuG). After completing his doctoral thesis in 2006, he continued studying Biomedical Technology at the RuG and obtained his MSc. degree (cum laude) in 2008. In November 2008, he started as PhD student at the Image Sciences Institute in Utrecht. In 2010 he moved to the Diagnostic Image Analysis Group (DIAG) in Nijmegen. The results of the work at DIAG and the Image Sciences Institute are described in this thesis.