UNIVERSITÀ DEGLI STUDI DI MILANO
Facoltà di Scienze Matematiche, Fisiche e Naturali
Dottorato in Informatica
XVII ciclo

A COMPUTER AIDED DIAGNOSIS SYSTEM FOR LUNG NODULES DETECTION IN POSTERO ANTERIOR CHEST RADIOGRAPHS

Tesi di Dottorato di
Elena Casiraghi

Supervisore: Prof.ssa Paola Campadelli
Coordinatore: Prof. Giovanni Degli Antoni

Anno Accademico 2003-2004

When I don't have any will to stand up in the morning "and start another game, another run", then I just need to think about those persons I'm thanking here. They are the proof that everyone of us can be an inestimable source of light, warmth and also new life.

This thesis has never been just mine: it is the thesis of and for My Parents and My sister and my Other Half; My Friends, the ones I still use to see and the ones I won't see anymore.

Aknowledgements

Thanks to Professor Paola Campadelli, for her inestimable support, both scientific and moral;
thanks to Doctor Angelo Vanzulli and Doctor Alberto Torresin from the Niguarda Hospital in Milan, for their kind cooperation and availability;
thanks to Giorgio Valentini for his good advices, and for his excellent collaboration to my reasearch;
thanks to Bram and all the people in the ISI group, for their precious comments and advices.

Contents

Introduction

Chest Radiography

X-rays were discovered by Conrad Wilhelm Rontgen in 1895 [68], while he was experimenting with electric current flowing in a cathode-ray tube. He noticed that the cathode emitted some unknown radiation that could be made visible with photographic paper and that passed through many materials. He did not realize that the radiation was a form of light and therefore called it "X-rays". His discovery heralded the age of modern physics and revolutionized the field of diagnostic medicine [11]. The great impact and change that this new technology caused in the case of chest examinations is proved by the fact that, since its discovery, chest radiographs make up 40% of all "X-Ray" examinations in total.
In Figure 1 a normal PA chest radiograph is shown. PA stands for Posterior-Anterior meaning that the radiation passes through the patient from back to front. The patient always faces the observer: the left side of the image shows the right lung.
The lung are radiolucent baskets of air, therefore they show up black in the image (by convention brightness indicates absorbed radiation). Within the lung fields, only bony structures and vessels are visible. The posterior ribs (in the back of the patient) are visible more clearly; it can be possible to follow them "turning" into the anterior ribs whose shadow appears less clearly. Other structures are visible, which may appear with very different shapes and brightness characteristics in other radiographs, or their appearance may vary a lot even if they are located in different positions in the same radiograph. This is one of the difficulties encountered by diagnosticians when looking at the thorax images. Some structures, such as gases or vessels, may seem similar to abnormal findings, since they do not have a unique representation. A specific example are the blood vessels: they may appear as round dots when running in the same direction of the X-Rays; in this case they could not be easily distinguished from abnormal findings such as lung nodules. This is just one example to motivate studies presented over the last two decades, such as [92] and [53], proving and also explaining why chest radiograph is one of the most challenging radiograph to produce technically and to interpret diagnostically. The authors of the mentioned

papers also noticed the large inter-observer and intra-observer differences occurring when radiologists rate the severity of abnormal findings.
Furthermore some studies by Samei and its colleagues [70] described and proved the great influence of noise in the detection of abnormalities by physicians. In the context of medical diagnosis from X-Ray images, "noise" is generally defined as the ensemble of all the variations (fluctuating intensities) present on the image, that interfere with the detection of the "true" signal being sought by the diagnostician, as for example lung tumors in the case of nodule detection. In this case the term "noise" is used to describe relative noise, or variations of the signal divided by the mean. On chest radiographs there are two major sources of such variations: "quantum noise" (mottle), which reflects the variations due to the finite number of x-ray quanta that form the image, and "anatomic noise", which reflects the highly "correlated" variations formed by the projection of anatomic features in the thorax, such as ribs, pulmonary vessels, and lung tissue. The influence of quantum noise in the analysis of chest radiographs to detect any kind of abnormality has been widely proved. This influence can be minimized either by increasing the patient dose, an action that is generally discouraged, or by improving the detecting quantum efficiency of the imaging system. Despite the substantial improvement in the detective quantum efficiency of imaging systems in the past few decades, which has reduced the level of quantum noise on chest radiographs, no or little improvement in the detection of abnormalities by radiologists has been assessed; this is a further proof of the difficulty of interpreting a radiographic image.
Nevertheless, chest radiographs is often used as the most common example of an "X-Ray Image"; this is due to the fact that, even though in the field of medical diagnosis a wide variety of more precise and more sensitive techniques are currently available, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), the chest radiography is still by far the most common type of procedure for the initial detection and diagnosis of every kind of abnormality, due to all the characteristics that make it the most easy and routinely available investigation procedure. They are for example non-invasivity characteristics, radiation dose and economic consideration.

Computer Aided Diagnosis Systems

Soon after the invention of the modern digital computer at the end of the $1940s$, a great deal of research work has been focused at the realization of computer systems performing tasks that had previously been carried out manually by humans. The basic motivations supporting these kind of researches were the speeding up of certain automatic and repetitive operations but also the support, and sometime even partial or entire substitution, of the human computation in the cases of difficult tasks requiring continuous

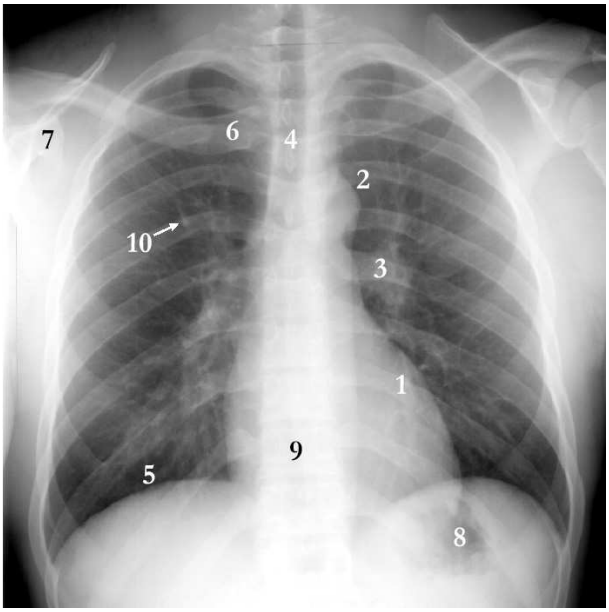

Figure 1: 1) The heart. 2) The aortic arch where the aorta bends. 3) The (in the case left) hilum, where the arteries and veins enter the lung. 4) A darker vertical stripe indicates the trachea. 5) Below the lung field the diaphragm starts. 6) Clavicle. 7) Shoulder-blade. 8) Usually stomach gases can be seen in the left diaphragm. 9) If there is enough contrast in the mediastinum (the area projected between the lung fields) the spine may be visible. 10)A round dot like this is the shadow of a vessel that runs in the same direction as the X-rays.

attention and hence subject to errors due to distractions. Given the possibility of the realization of such systems, their advantages were of course the better performances in terms of time (since computers process certain kind of data much faster then the human being), costs (1 computer program can substitute the work of more persons), and error rate (the attention of the computer never decreases). In the case of medical image analysis on thorax radiographs the first articles introducing the idea of such systems appeared in the 1960*s* ([7], [59]). Duncan and Ayache in [20] and then Ginneken in [86] presented good reviews of all the work presented in this field; their description starts from the 1970*s*, when more specific articles describing techniques and studies specifically designed for computerized detection of abnormalities in chest radiographs began to appear. These early studies were referred in a review by Conners et *al.* in [16] as attempts to *fully automatize the chest exams*, and displayed a considerable optimism regarding the capabilities of computers to generate complete diagnoses. Over the decades this expectation has subsided, as seems to have been the case with the early enthusiasm regarding artificial intelligence in general. Instead of trying to substitute diagnosticians, the current trend is to create Computer

Aided Diagnosis (CAD) Systems. They may be defined as a "diagnosis made by a radiologist who takes into account the result of a computer analysis". Thus, the aim of the computer program is to process the image in order to extract and create information useful for physicists and radiologists during their decision making process. In the field of computer aided diagnostic, almost all work is applied to frontal (Postero-Anterior) chest radiographs. Few studies are aimed at lateral radiographs, as for example the one presented by Armato et *al.* in [3] and by Carrascal et *al.* in [14]. The main purposes of the documented automatic systems for PA chest radiographs analysis are well summarized by Ginneken in his review; they are divided into three main areas listed below:

Segmentation
- Lung field Segmentation
- Rib cage
- Other structures such as clavicles

General processing
- Enhancement
- Subtraction techniques

Analysis
- Size measurement (such as cardiothoracic ratio, estimation of total lung volume)
- Lung Nodules Detection
- Texture Analysis

The first two areas basically include applications that are mandatory for the analysis of any kind of abnormalities in the chest radiographs. Their purpose is indeed to perform the first two steps of the normal process carried out by the diagnosticians: first, they extract the area of interest where they will be searching for the pathology (i.e. they focus on the lung fields), then they try to look better at it (this can be done via computer by means of an enhancement of the image designed specifically for the required diagnosis). The last area includes all the programs simulating the analysis method performed by humans; it is based on steps for knowledge extraction and its interpretation to produce a final diagnosis.

Lung Nodule detection in Postero anterior chest radiographs

Lung cancer is a major cause of mortality in the Western World; clinical reports state that in the United States it is one of the leading cause of death, second only to cardiovascular diseases. In 1999 the American Cancer Society reported that an estimated 171600 new cases of lung tumors where discovered in the United States, which accounted for 14% of cancer diagnoses, and an estimated that 158900 deaths were caused, accounting for 28% of all cancer deaths (ACS, 1999). American Lung Cancer Society statistics indicate that the 5-year survival rate for patients with lung cancer can be improved from an average of 14% to up to the 49% if the disease is diagnosed and treated at its early stage ([61]). The treatment options are surgery, radiation therapy and chemotherapy. For the case of localized but already large cell lung cancer, surgery combined with radiation therapy and chemotherapy is usually the treatment choice, even though no guarantee of success in the treatment can be given to the patients. In the case of small cell lung cancer, patients are treated with chemotherapy alone or combined with radiation.

The correct detection and diagnosis of solitary, circumscribed pulmonary nodules in radiographic chest images are of great importance because many of these lesions are early primary bronchogenic carcinomas, classified by the American Joint Committee on Cancer Staging as "AJC stage 1" lung cancer. Several studies, reported in [87], [75], [10] and [33], proved that the survival rate can be improved greatly if the carcinoma is removed at this stage. Moreover failure to detect these lesions may results in costly delays in appropriate treatment. At their first stage the tumors are generally referred as nodules for they are generally bright round shaped abnormalities.

The previously cited numbers and reports explain why the detection of subtle and early staged lung nodules on chest radiographs is one of the outstanding challenges in the field of medical diagnosis on chest radiographs. In the 1980*s* Stitik and his colleagues found that a single radiologist did miss 32% of all lung nodules viewed retrospectively, and that two radiologists working together missed only the 15% ([29], [60], [76]). The same average miss rate was calculated during several contemporaneous studies, such as the ones presented by Forrest and Friedman in [23], and by Austin in [6]. The most impressive number has been reported again by Muhm et *al.* in [60], who conduced a long-term project on lung cancer detection at the Mayo Clinic to discover that 90% of peripheral lung cancers were visible in radiographs produced earlier than the date of the carcinoma discovery by the radiologist. Despite many technologic advances in the last 4 decades, this numbers and statistics have not been improved, as reported by Gavelli and Giampalma in [25].

The most common techniques for lung tumors detection currently used and known include chest radiography, cytologic analysis of sputum samples, fiber-optic examination of bronchial airways, and finally Computerized Tomography scans (CT) and Magnetic Resonance Imaging (MRI). Although the last two are more sensitive and precise techniques, chest radiology remains the initial and most common procedure since it is the most cost-effective, the most routinely available, and also the most dose effective diagnostic tool for the detection of lung cancer, as reported by Murphy in [61]. It is often preferred also for its ability of revealing some unsuspected pathologic alterations (lung nodules are the most cited example). Freedman and his colleagues in [101] estimated that the 90% of lung cancer detection now takes place on chest radiographs. All these clinical reports explain why the discovery of the early stage nodules in radiographic images is a source of major concern in the field of medical diagnosis for physicians, radiologists and also patients. Several studies in the last decade have been devoted at the analyzing the major causes of errors by the radiologists; some of them, as for example the ones presented by Samei and his colleagues in [70] and [71], try to asses the influence on the detection accuracy of the noise in the image, proving that both anatomical noise and the one created by the imaging device, named as quantum noise, have a great influence in the detection accuracy. Giger (in [26]) gave a list of the observer error which may cause several lesions to be missed. They are generally due to the camouflaging effect of the surrounding anatomic background on the nodule of interest, or to the quantum noise caused by the imaging system, or to the subjective and varying decision criteria used by the radiologists. Under-reading of a radiograph may also be due to a lack of experience, lack of clinical data, a premature discontinuation of the radiograph reading because of a definite finding, focusing of attention on another abnormality by virtue of a specifical clinical question, failure to review previous radiographs, distractions, and "illusory visual experiences" ([73], [2]).

A computer approach is attractive because it has the potential to provide objective and consistent results; thus, an automatic scheme that alerts radiologists to the location of highly suspected lung nodules should allow the number of *False Negative* (or missed nodules, see Appendix C) diagnosis to be reduced. This could lead to earlier detection of primary lung cancer and of metastatic nodules and, it is hoped, a better prognosis for the patient. Moreover the advent of digital thorax units and digital radiology departments with Picture Archiving Communication Systems (PACS) makes it possible to use computerized methods for the analysis of chest radiographs as a routine basis. A successful detection scheme could eventually be hardwire implemented for on-line screening of all chest radiographs, prior to viewing by a physician. Thus, chest radiographs ordered for medical reasons other than suspected lung cancer would also undergo careful screening for nodule detection. The motivation of this last sentence is that there is

no national screening program currently in place for lung cancer, so that early stage lung cancer are often discovered in the form of solitary lung nodules when a chest radiograph is obtained from a patient for another purpose ([71]). The feasibility of Computer Aided System for lung nodule detection was demonstrated by several authors such as Kobayashi et *al.* in [79], Matsumoto et *al.* in [80], MacMahon et *al.* in [54]; they proved the possibility of improving radiologists' detection accuracy for lung nodules in chest radiographs by means of a system providing the radiologists with a "second opinion".

These reasons, together with the fact that a shortage of radiologists has been predicted in the next decades, explain why in the last two decades lot of research work has been focused on the creation of CAD systems aimed at lung nodules detection. Kundel and Revesz in [45] introduced the concept of *conspicuity* to describe those properties of an abnormality and its surrounding which either contribute to or distract from its visibility. Many other studies were concentrated on the factors that influence the visibility of the nodules, and found out that the visibility of nodules of same size varies according to their location in the same radiograph (an example are the peripheral nodules, that are more difficult to be detected for their position); on the other hand also the size influences the visibility of nodules in the same location and in the same radiographs. Thus, a computerized search scheme would have to be capable of locating nodules of different size and with varying degrees of conspicuity, i.e. nodules immersed in backgrounds of various anatomic complexity.

Almost all the methods presented in the literature use a three step approach for nodule detection, that resembles the normal process of the human intelligence. In the first step the lung field is segmented and the area of the lung is processed to produce an enhancement of the image with the specific aim of increasing the visibility of the nodules; in the second step all the regions that may contain nodules are extracted; they are referred as *False Positive regions* when they don't contain nodules in reality, and *True Positive regions* if they are really containing nodules; the final step consists of eliminating as many *False Positives* as possible without sacrificing too many *True Positives*. Even though a great deal of research work has been focused on Computer Aided Systems for lung cancer detection and a wide variety of them has been already proposed, none of them has been applied successfully on clinical trials. At the state of the art, the results presented are still far from being useful in practice since too many *True Positives* are lost in the last step or too many *False Positive regions* are kept. Ginneken in his review ([86]) made a schematic list of all the methods presented. From his table it is clear that there is a considerable overlap between the methods employed in various studies, since they are often presented by the same researchers; despite the big number of studies reported, the number of groups that have been working, or are working on this problem is indeed

limited. Lately this number is even decreasing due to the advent of new and more appealing techniques (such as CT or MRI images), which may seem attractive because the development of computerized systems to detect tumors on the chest scans, produced by them, is surely easier. Nevertheless, as pointed out in the previous section, they are not so routinely available, cost and dose effective: chest exams using them are generally required by diagnosticians when they already doubt the presence of a tumor or when it has been already detected, to check its condition; in this cases radiologists do not need further help to analyze the images obtained, since they are so detailed and precise representation of the chest that no error can be done and no information could be missed. This is the reason why a computerized system aimed at early stage lung nodules detection and using as input CT or MRI images is considered to be of less use.
Another problem regarding CAD systems working on radiographs is that almost all the results presented do not prove the efficiency of the methods: most of the systems are tested on private and very little databases; other papers report results on radiographic images where a simulated nodule has been superimposed. This raises another problem of creating common databases that can be used by different researchers to test and compare their methods. One database, described in the section "Materials", has been created by the Japanase Society of Radiological Technology (JSRT) ([72]), meanwhile another is being actually created in the AZU hospital in Utrecht where a chest screening program has been started at the beginning of this year.
Once a robust method for nodule detection will be realized the next step will be the development of systems aimed at the classification of the detected nodules as benign or malignant. Few studies have appeared on this subject where the nodule regions are still extracted manually. They are presented by Nakamura et *al.* in [62], and by Gurney and Swensen in [30].

Outline of the work

This thesis describes a Computer Aided System aimed at lung nodules detection. The fully automatized method developed to search for nodules is composed by four steps. They are the segmentation of the lung field, the enhancement of the image, the extraction of the candidate regions, and the selection between them of the regions with the highest chance to be True Positives. The steps of segmentation, enhancement and candidates extraction are based on multi-scale analysis. The common assumption underlying their development is that the signal representing the details to be detected by each of them (lung borders or nodule regions) is composed by a mixture of more simple signals belonging to different scales and level of details.
The last step of candidate region classification is the most complicate; its

task is to discern among a high number of candidate regions, the few True Positives. To this aim several features and different classifiers have been investigated.

In Chapter 1 the segmentation algorithm is described; the algorithm has been tested on the images of two different databases, the $JSRT$ and the $Niguarda$ database, both described in the next section, for a total of 409 images. We compared the results obtained with another method presented in the literature and described by Ginneken, in [85], as the one obtaining the best performance at the state of the art; it has been tested on the same images of the $JSRT$ database. No errors have been detected in the results obtained by our method, meanwhile the one previously mentioned produced an overall number of error equal to 50. Also the results obtained on the images of the $Niguarda$ database confirmed the efficacy of the system realized, allowing us to say that this is the best method presented so far in the literature. This sentence is based also on the fact that this is the only system tested on such an amount of images, and they are belonging to two different databases.

Chapter 2 is aimed at the description of the multi-scale enhancement and the extraction methods.
The enhancement allows to produce an image where the "conspicuity" of nodules is increased, so that nodules of different sizes and located in parts of the lungs characterized by completely different anatomic noise are more visible. Based on the same assumption the candidates extraction procedure, described in the same chapter, employs a multi-scale method to detect all the nodules of different sizes. Also this step has been compared with two methods ([8] and [1]) described in the literature and tested on the same images. Our implementation of the first one of them ([8]) produced really poor results; the second one obtained a sensitivity ratio (See Appendix C for its definition) equal to 86%. The considerably better performance of our method is proved by the fact that the sensitivity ratio we obtained is much higher (it is equal to 97%) and also the number of False positives detected is much less.

The experiments aimed at the classification of the candidates are described in chapter 3; both a rule based technique and 2 learning systems, the Multi Layer Perceptron (MLP) and the Support Vector Machine (SVM), have been investigated. Their input is a set of 16 features. The rule based system obtained the best performance: the cardinality of the set of candidates left is highly reduced without lowering the sensitivity of the system, since no True Positive region is lost. It can be added that this performance is much better than the one of the system used by Ginneken and Schilam in [1], since its sensitivity is lower (equal to 77%) and the number of *False Positive* left is comparable. The drawback of a rule based system is the need of setting the

thresholds used by the rules; since they are experimentally set the system is dependent on the images used to develop it. Therefore it may happen that, on different databases, the performance could not be so good.
The result of the $MLPs$ and of the $SVMs$ are described in detail and the ROC analysis is also reported, regarding the experiments performed with the SVMs.
Furthermore, the attempt to improve the performance of the classification leaded to other experiments employing SVMs trained with more complicate feature sets. The results obtained, since not better than the previous, showed the need of a proper selection of the features. Future works will then be focused at testing other sets of features, and their combination obtained by means of proper feature selection techniques.

Materials

All the method presented in this thesis work have been developed and tested on a standard database acquired by the Japanese Society of Radiological Technology. It is a standard database containing a total of 247 radiographs: 154 containing lung nodules of different diameter (ranging from 5 to $35mm$) and subtleties (ranging from 1 to 5, from "obvious" to "extremely subtle"), and 93 of patients with no disease. The images were digitized with a $0.165mm$ pixel size, a matrix size of 2048 by 2048, and 4096 grey levels. This database contains images with different characteristic of brightness; moreover some images may be affected by noise due to the imaging system (dark lines, or bright borders). The position of the chest is generally vertical, but some images are radiographs of patients with chest abnormalities, in some other one of the two lungs is not visible due to the inspiration of the patient. Other type of noise may be due to external devices that appear as external structures contained in the chest. Examples of these images are shown in Figure 2. When present, the tumors are indicated by the intersection of vertical and horizontal lines.
A detailed description and an article describing the procedure used to create the database has been reported in [72].

The segmentation algorithm has been tested also on 162 radiographic images acquired in the Department of Radiology of the *Niguarda* Hospital in Milan. The images were digitized with a 0.160 mm pixel size, a maximum matrix size of 2128 by 2584, and 4096 grey levels. They have not a fixed squared size and their resolution is not the same as the ones in the $JSRT$ database, since they have been created with a different imaging system. Their characteristics are completely different from the ones in the $JSRT$ Database; they are affected by a completely different type of noise due to

the imaging system: it is often more visible and appears as more scattered; moreover lots of images contain chests with structural abnormalities. The patients are often tilted or rotated, the position of the chest in the radiograph is not always at the center and there is a lot of variation in the size of the thoraxes. Some of these images are visible in Figure 3.
For the *Niguarda* Database no information regarding the presence of lung nodules is available. For this reason they have been used as input only to the segmentation algorithm.

A further observation that must be added is that the big differences between thorax images, both in the same database and especially in different databases, make the development of systems aimed at chest analysis even more challenging. This is for the impossibility to define a general model describing the physical structure of the chest or the noise affecting the radiographs.

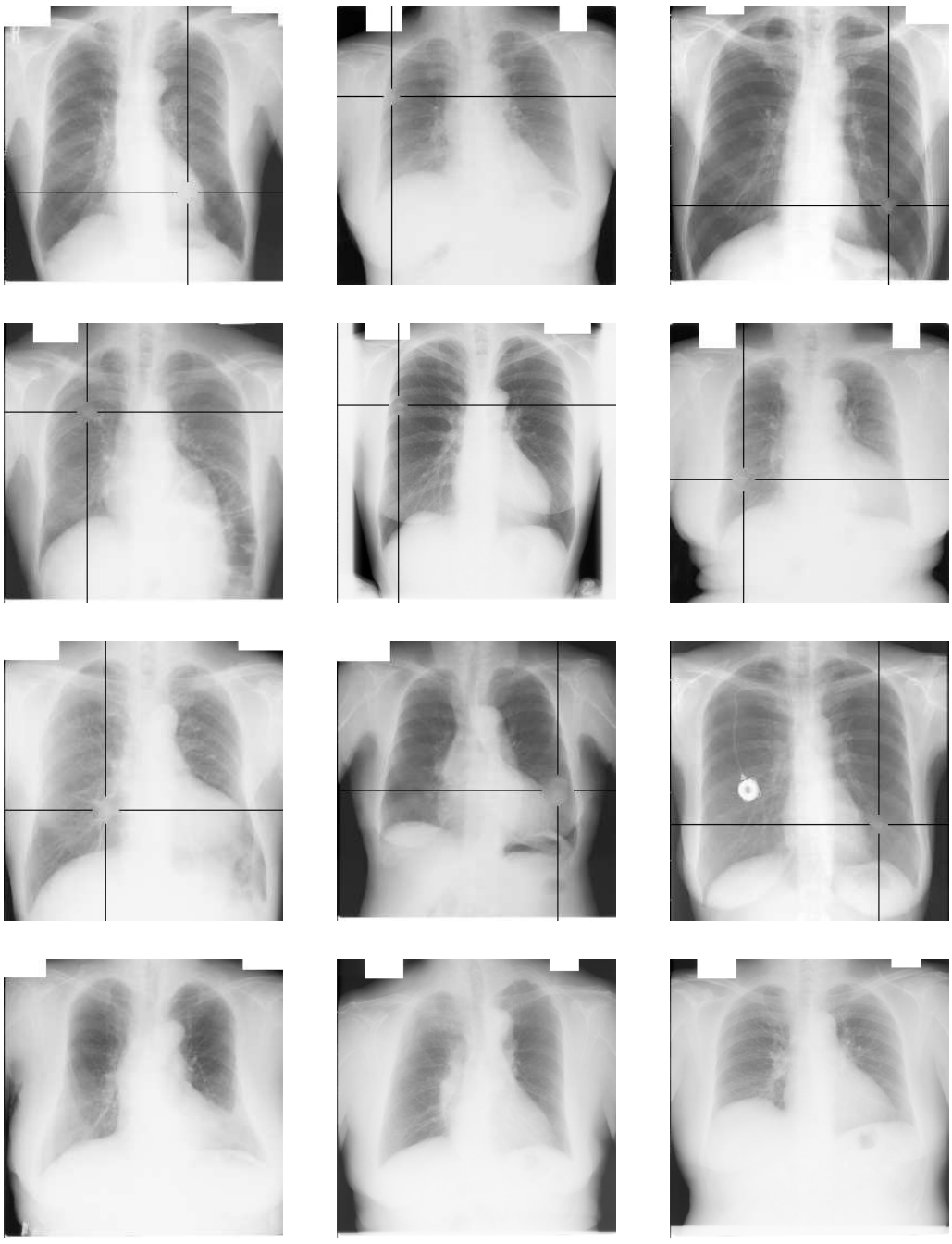

Figure 2: *examples of chest radiographs in the JSRT Database*

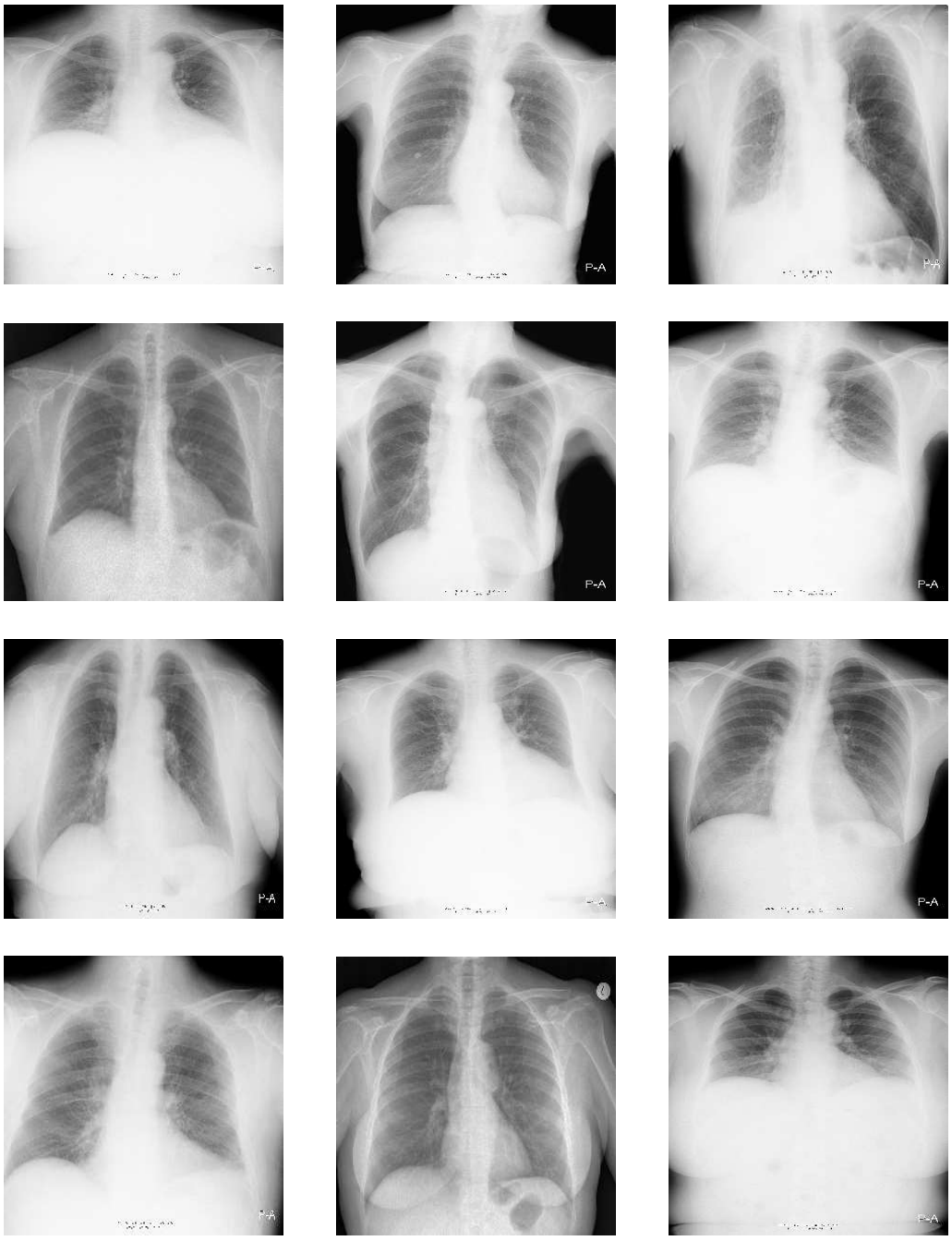

Figure 3: *examples of chest radiographs in the Niguarda Database*

Chapter 1

Segmentation of the Lung Field in Digital Postero Anterior Chest Radiographs

The first step for the development of an automatic system aimed at any kind of computerized analysis on digital chest radiographs, is the automatic segmentation of the thorax image in order to extract the area of the lungs. In this chapter a segmentation method is described, whose result is a close contour strictly enclosing the lung area. The method has been developed and tested on the standard *JSRT* database, containing 247 radiographs. To have a further validation of its efficacy it has been tested also on 162 images acquired from the *Niguarda* hospital in Milan. The results obtained and the comparison with other methods presented in the literature, prove the robustness of the algorithm developed.

1.1 Introduction

The reason motivating the widespread use of chest radiography, for the initial detection of every type of abnormal pulmonary condition, is the fact that it is by far the most routinely available type of screening procedure. Between all the other more precise and complicated procedures this is indeed the one preferred by physicians for its ability of revealing some unsuspected pathologic alteration, for its non-invasivity characteristics, radiation dose and economic considerations. The segmentation of the lung field is surely the first and mandatory step of an automated system aimed to any type of computer analysis on chest radiographs, so that the algorithms for the identification of the abnormalities will be applied just to the lung area defined. This is the reason why a great deal of research work has been focused on the segmentation of the lung field in Postero-Anterior chest radiographs.
Several theoretical studies in the past decades have been focused on the

demonstration of the fact that the automatic segmentation of the lung fields is indeed a hard problem from a computer vision point, and several reasons have been highlighted. Studies by Vyborny [92] and MacMahon [53] are aimed at explaining why chest radiograph is one of the most challenging radiograph to produce technically and to interpret diagnostically. First, there are large anatomical differences between the chests of different patients. Second, the habitus, the position and finally the level of inspiration of the person in the moment of the acquisition of the radiographs can cause big variations in radiographs of the same patient; the inspiration especially has a great impact on the visibility of some parts of the lung. Third, the setting of the chest unit, particularly the peak tube voltage, determines the visibility of several structures such as the bones and the vessels ([5]). One thing to be noticed is that quantum noise, due to the imaging system, is not considered as deleterious and it may not affect the performances of a segmentation method; it is proved that the disturb it causes in the images influences the visibility of certain subtle pathologies, meanwhile it doesn't decrease the visibility of the rib cage. The last but most important cause of problems is that radiographs are projection images and hence represent all at once superimposed structures. The task of segmentation is then mislead if some external structures are superimposed to the real borders of the lungs. Examples are the hilum, where the arteries and veins enter the lungs, bony structures such as the clavicles, the spine and the shoulder-blade, that are often better visible then the rib cage boundaries, the heart, the diaphragm, that is often hiding the costophrenic angles, and the intestinal gases in the left lung area. It is obvious to understand that the incorporation of knowledge is strictly needed to recognize all the listed problems and solve, or even avoid, the errors caused by their effects.

Another important issue about lung segmentation is the definition of the lung area to be detected and delineated. All the methods presented in the literature are indeed aimed at the definition of the most visible parts of the left and right lung, i.e. the ones which are not hidden behind the diaphragm, the heart and the spinal column. This kind of segmentation identifies an area, which will be referred as *visible lung area*, that is optimal when the abnormalities or pathologies to be detected by the computerized system, such as tuberculosis, are restricted just to the visible lung area or, when present, are spread all over the lung. In these cases it is better to restrict the processing just to this search area since it is much less extended hence needing less computational costs. This *visible area* is not enough for lung nodules detection systems, since tumors may be present also in the *not visible parts* of the lungs, and in these cases they would be lost by the system. Examples of the visible lung area extracted by the algorithm presented by Ginneken are shown in Figure 1.1; they are thorax images of patients with lung tumors. Note that some nodules are outside the segmented area.

Another weak point of several methods presented in the literature is the fact

that they are based on assumptions regarding the position, orientation and size of the chest in the image.

At the state of the art two main approaches have been presented in the

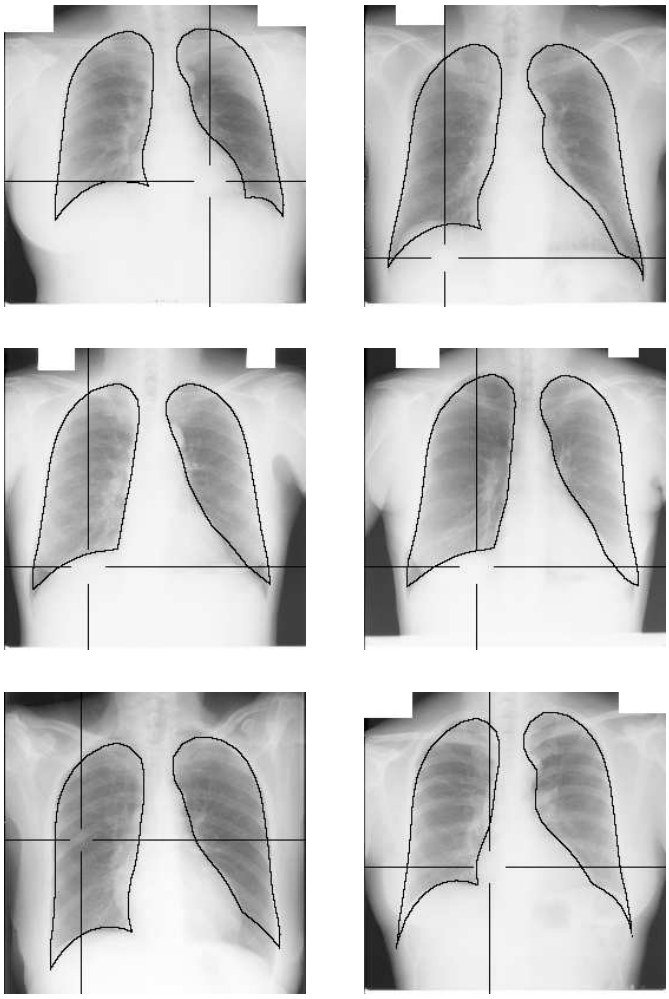

Figure 1.1: General result of a lung segmentation algorithm and the position of the nodules

literature for solving the task of chest radiographs segmentation: they are rule-based reasoning and pixel classification.

We classify as rule based systems those composed of algorithms performing consecutive steps, and each containing specific processing and usually certain adjustable parameters. Examples of such schemes have been presented by Xu et *al.* in [96] and [97], Duryea and Boone in [21], Carrascal in [14], Armato et *al.* in [4], Pietka in [65], Chen et *al.* in [15], Brown et *al.* in [9].

Pixel classification methods are based on the classification of the pixels of the image into two classes, *"lung pixels"* and *"not lung pixels"*; the classifi-

cation is based on a set of features calculated for each pixel itself. Examples of such systems using Neural Networks for learning and classification have been proposed by McNitt-Gray in [58] and [57], by Hasegawa et *al.* in [32], and Tsujii in [81]. A less common pixel classifier has been experimented by Vittitoe and described in [90] and [91]: the author developed a pixel classifier for the identification of the lung area using Markov random field modelling.
Other approaches, but less common, identify the lung area making use of classical methods for image segmentation, such as the Active Shape Models, introduced by several authors as for examples Jain (in [36]) and Cootes (in [17]) who experimented them also for lung segmentation in chest radiographs. Ginneken presented an extension of these models for image segmentation in general, and then described its application also for the specific task of lung segmentation (see the work described in [85] and Figure 1.1 for some examples of the results obtained). Another possible approach for the detection of lung borders is the registration and matching of a common reference image to the input radiograph. Despite this is a common scheme for most type of medical image segmentation problems, for the analysis of chest radiographs its use has been generally restricted to the step of ribs detection and it has not been explored for the task of lung definition.
An interesting study was described in [84] by Ginneken, who compared the performances of 9 simple algorithms based on different techniques. The first 8 of them are simple methods employing matching, rule based reasoning and pixel classification based on different features (such as intensity or location); the ninth is a hybrid system composed by the proper integration of a rule based reasoning technique and pixel classifiers of various type. The best performances were obtained with this last system; the results obtained with the systems using rule based reasoning and pixel classifiers were less good but still promising. This comparison is helpful to explain the widespread use of the techniques just mentioned, and prove that knowledge based processing is necessary to solve the task of lung area segmentation.

1.2 Purpose and Materials

In this chapter we describe our segmentation method which identifies the lung area in digital Postero-Anterior (PA) chest radiographs. This is the first step of an automatic system for the detection of lung nodules; this motivates the decision to include into the segmented lung field also the area behind the heart, the spine and the diaphragm, where lung nodules may still be present. As pointed out in the previous section this choice differs from that of all the other methods presented in the literature. Another difference from these methods is that the algorithm avoids all kind of assumptions

about the position, orientation and size of the thoraxes. It works with images where the chest is not always located in the central part of the image, it can be tilted and it can have structural abnormalities.
When the whole area of the lungs has been detected a further processing is aimed at the separation of the parts hidden, behind the heart, the diaphragm or the spinal column, from the visible ones. This step is helpful to increase the information about the lung field produced by the segmentation. This kind of information is indeed required to get a better description about the "nodule candidate regions" extracted by the automatized algorithm that will be described in the next chapter.
To detect the lung borders the method employs two different edge detection techniques and then combines their results. The first one is based on the application of first derivatives of gaussian filters taken at 4 different orientations (See Appendix A for a description of the oriented derivative operators). The second one is a multi-scale method based on the application of the Laplacian of Gaussian (LoG) operator (See Appendix B for a description of this operator): it employs 3 different scales to search for all the structures of the lung border. This multi-scale analysis constitutes the great difference with all the methods presented in the literature and it is based on the inherent multi-scale property of the lungs. Specifically, the lung borders are formed by structures which differ for their level of detail and their dimensions; some of them can be identified when analyzing the image at a finer scale, others are detected at coarser scales. A good introduction and motivation of the utmost importance of multi-scale methods for image analysis and information retrieval in general can be found in [48], [49] and [50].
To separate the visible and not visible parts of the segmented lung area a classical clustering method presented by Arbib and Uchiyama in [82] and based on competitive learning has been used; the refinement of the area thus detected is carried out by its integration with the result of simple derivative filters to detect the edges in the image. This method proved to be helpful also to detect and correct some errors that couldn't be avoided by the first step of segmentation to define the full lung area.

Materials

The whole method has been developed and tested on the radiographs of the standard database acquired by the Japanese Society of Radiological Technology. For the first segmentation part, the images have been down-sampled to a dimension of 256×256 pixels, for the second part of segmentation their size has been reduced to 512×512 pixels: these sizes have been experimentally chosen in order to reduce the computational costs of the algorithm

without losing any details that could have any influence on the performances of the overall segmentation algorithm. These down-sampled images will be referred as the *Original images*.
To have a further check on the efficacy of the method the algorithm has been tested also on 162 radiographs acquired in the Department of Radiology of the *Niguarda* Hospital in Milan. Note that even though these images are not squared and their resolution is not the same as the ones in the *JSRT* database, they have been down-sampled to the same dimension of 256×256 (and 512×512 for the second part) without causing any problem to the algorithm, as proved by the results obtained.
In the following description the lungs will be referred as "left" and "right" with respect to their position in the image. The coordinate system used to give the position of the pixels in the image has the origin in the top left corner, the positive Y axis corresponding to the height and the positive X axis to the width.

Previous work

The algorithm described in the rest of the chapter is the result of a study focused on lung segmentation. During this study several classical techniques for image segmentation have been experimented, and their performances, when dealing with thorax images, have been assessed. Examples of such techniques are the diffusion algorithm presented by Perona in [64], the Watersheed algorithm presented in [89], fuzzy classification techniques, and iterative thresholding methods. The experiments executed with the first two techniques have proved their inadequacy for the task of lung field detection, meanwhile better results have been obtained with the last two. The two systems realized and based on these two techniques have been described and compared in [74]. Moreover their combination with the classical edge detectors techniques has lead to an algorithm showing good performance when tested on the *Niguarda* Database; it is described with details in [12] and some results are shown in Figure 1.2.
Despite the good results obtained on the *Niguarda* database, on the *JSRT* database the number of errors was high (around 40). This motivated the development of the algorithm described in the rest of this chapter.

1.3 Segmentation of the full lung area

1.3.1 Edges detected by derivatives of gaussian filters

A rough outline of the lung borders is obtained by searching the edge pixels at four different directions, 90°, 180°, 45° and 135° since they correspond to the orientation of most of the edges in the lung contour. Their location is obtained by means of steerable first derivatives of gaussian filters (See Ap-

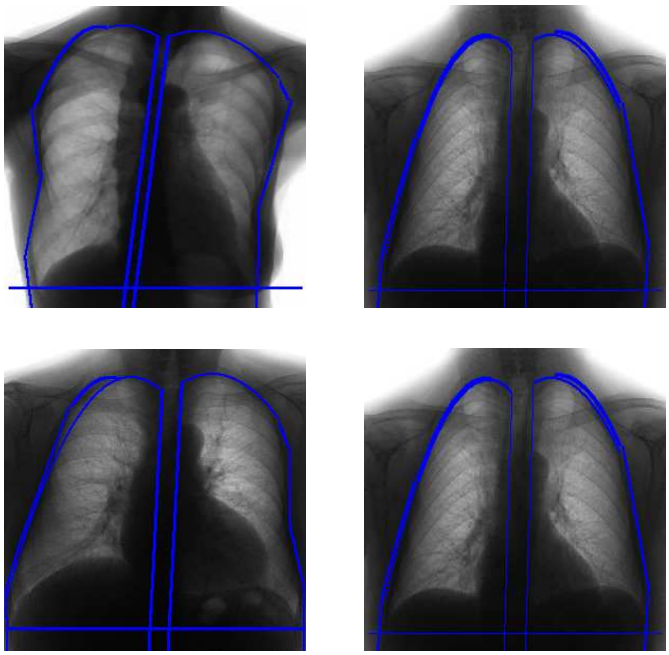

Figure 1.2: some results of the segmentation described in [12]

pendix A) at a scale $\sigma = 1$, oriented along the four directions perpendicular to the ones previously mentioned. Maintaining for each direction the 10% of the pixels with the highest gradient value four different *Binary edge images* are created, B_{90}, B_{45}, B_{135},B_{180} (the subscript index is referring to the main orientation of the borders detected), whose union is a binary image, called *Global edge image*, containing a rough outline of the lung borders but also spurious details (see Figure 1.3). Those that are located along the margins of the image are mainly due to noise caused by the acquisition system. To eliminate them the algorithm analyzes the connected regions located in the vertical margins and with a width of at most 20 pixels, and the ones located in the horizontal margins and with a height of at most 10 pixels. They are discarded if their *rectangularity* is bigger than 0.7, where the *rectangularity* of a region is defined as the fraction of the area of the region itself and the area of the maximum bounding box including it.
Two examples of the cleaned *Global edge image* are shown in the left of Figure 1.4; Other unwanted details in the *Global edge image* are due to bony structures external to the lung field (such as the clavicles, the collar bones, the vertebrae). Some of them are discarded by eliminating from the image all the connected regions whose area is less then 0.05% of the image area. Note that in this way also some regions that are part of the lung borders may be deleted; to be able to recover from possible errors, the discarded regions are stored in an image called *Little regions image*: they will be recuperated

later, if needed. Since the edges belonging to bony structures attached to the lung field, such as the clavicles or the collarbones, are basically oriented at 180°, they are mainly contained in B_{180}. Therefore a new image is built, called *Partial edge image*, that does not contain them; it is the union of B_{90}, B_{135} and B_{45} cleaned in the same way previously described; in Figure 1.4 the *Partial edge image* is shown, before (central column) and after cleaning it (right column).

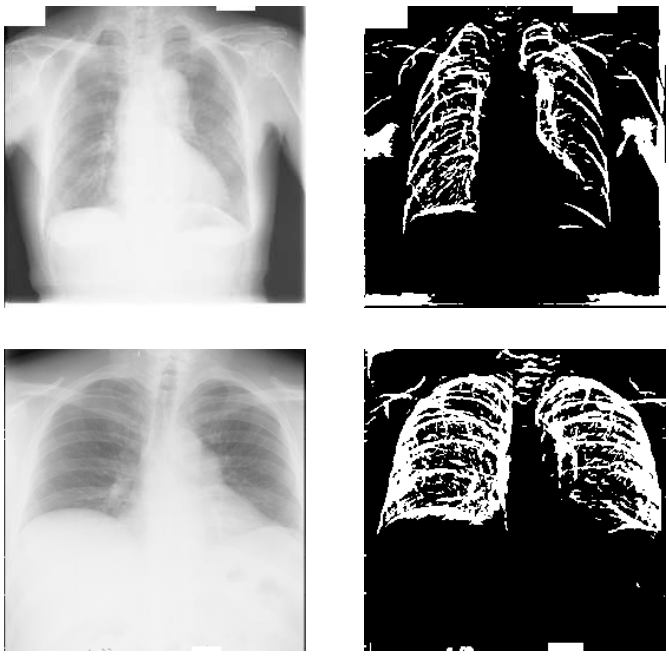

Figure 1.3: *Original image - Global edge image*

1.3.2 Finding the axis and further cleaning of spurious details

The *Partial edge image* is used to find the axis of the lung field. For each horizontal line in it, the pixel in the center of the segment connecting the leftmost and rightmost pixels is found and its position is stored. A polynomial fitting method that minimizes the $\chi - square$ error statistic, is used to find the axis of the lung field. In Figure 1.4 the points used are shown red colored in the *Partial edge image*, together with the axis found. It is always located in the center of the dorsal column and between the two lungs. This fact allows to delete those edges in the *Global edge image* belonging to the dorsal column or to the neck. To this aim a "stripe" around the axis is defined (it is containing the axis and has a width that is 1/20 of the width of the image) and all the connected regions in it are deleted. The *First edge image* thus obtained (see the right column in Figure 1.4) is a first outline of

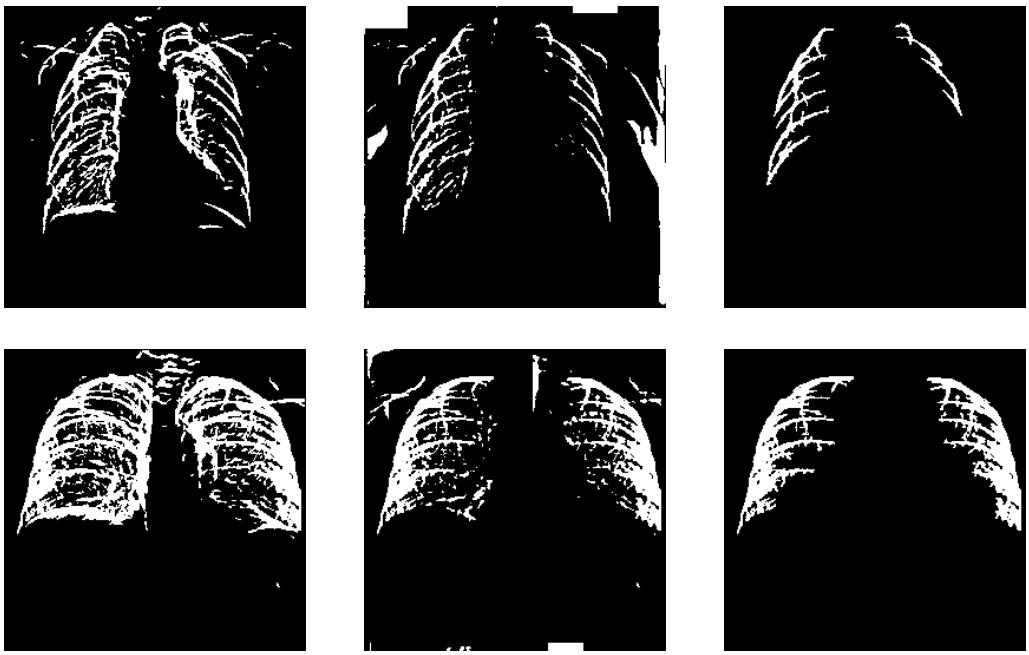

Figure 1.4: Cleaned *Global edge image* - *Partial edge image* created and then cleaned

the lung borders. This image still contains errors such as missing borders and bony structures external to the lung field.

1.3.3 Contour finding by tracking the edges detected with LoG operators

In this section our method based on edge detection by Laplacian of Gaussian (LoG) operators (See Appendix B) at 3 different scales, followed by an edge tracking technique is described. As already said, the choice of using a multi-scale approach is well motivated by the fact that the borders of the lung fields are composed by distinct structures and details visible at different scales ([48]). The result of the integration of the information detected by the three filters is the detection of a continuous path along the lateral and top borders of the lung. The contour found will then be combined with the *First edge image* to get a precise segmentation of the lung field. The method used works separately on the left and right lung. Here, it will be described just for the left side since the steps applied to the other lung are basically the same (left and right lung are distinguished by means of the axis found).

The first step of the technique described in this section, is the definition of three different binary images, L_1, L_2 and L_3, created by simply applying to the *Original image* the LoG filter at three different scales, ($\sigma = 1, 2, 3$), and

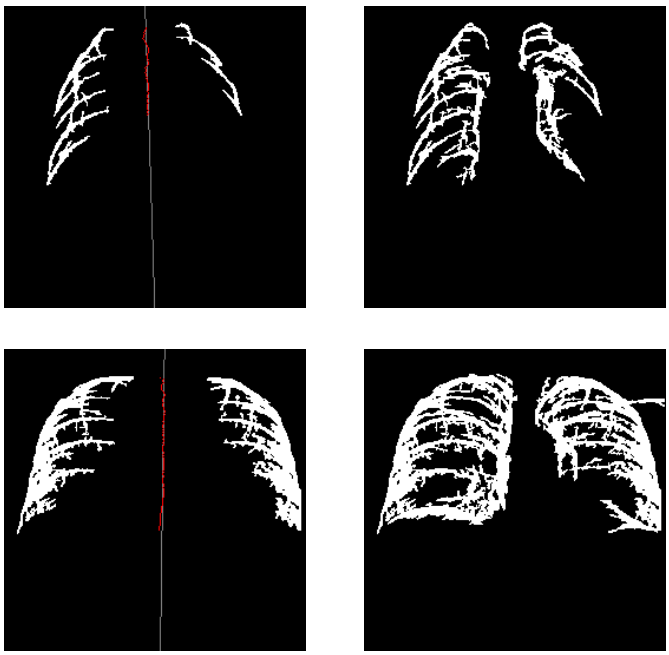

Figure 1.5: *Global edge image - First edge image*

selecting the pixels with positive values. It can be noticed from Figure 1.6 that L_1, L_2 and L_3 capture details of the chest radiograph visible at different scales. Their pixels are analyzed by an edge tracking algorithm to create three possible contours, one in each image; these will then be combined to get a final one.

Settings for the edge tracking procedure

The first step to start the edge tracking algorithm on the left border of the L_σ ($\sigma = 1, 2, 3$) images, is the selection of its starting point $P_{left} = P_{left}(x, y)$, located on the rib cage boundaries. To find it the intersection between the *First edge image* and the *Partial edge image* is considered. This choice is due to the fact that this intersection does not contain the errors due to the horizontal bony structures, such as the clavicles, meanwhile it still contains the most visible horizontal and vertical edges on the top part of the lung. P_{Left} is found by taking the point in this image that is the nearest to a line oriented at 135°, and passing trough the origin of the coordinate system used. If more than one point is found the topmost is kept. This is a very simple method to approximately find the point where the change of the curvature of the lung borders is maximum; indeed, this is the point where the edges orientation changes from horizontal (on the top border) to vertical (on the lateral border).
Testing this searching method on the overall set of images in the database

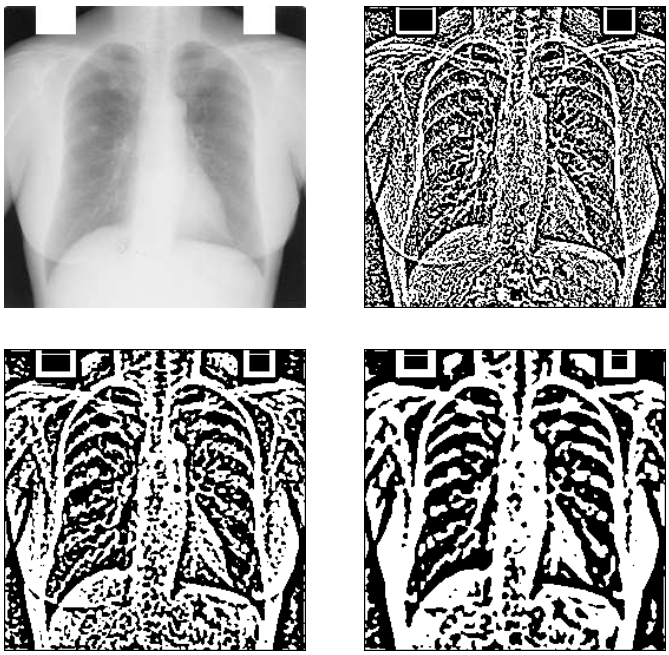

Figure 1.6: *Original image* (top left)- Positive pixels in the Laplacian images $L_\sigma(\sigma = 1, 2, 3)$

a proper starting point has always been detected; this is generally located on the lateral part of the lung boundaries and near to the top. In the next paragraph the edge tracking algorithm is described. Its input is the starting point P_{Left} and a binary image, in this case L_σ ($\sigma = 1, 2, 3$), on which the path is to be searched; the output is a continuous path running approximately from the top to the bottom of the lung field.

Edge tracking algorithm

The algorithm is composed by two steps, each one detecting a continuous contour, *Bpath* and *Tpath*, along the borders of the lung field. *Bpath* is the one created from the point P_{Left} to the bottom of the lung fields. *Tpath* runs from P_{Left} to the top of the lung fields.

Construction of the path running to the Bottom

In this sub section the procedure used to define the *Bpath* is described. The first thing needed is to check if the pixel in position P_{Left} is set to 1 in the input image L_σ [1]. If it happens this will be used as starting pixel P_S^B

[1]Note that this point was chosen from the intersection between the *First edge image* and the *Partial edge image*, hence it may not be similarly set to 1 in the binary images, L_1, L_2 and L_3, used as input.

for the descending path. If this is not the case another point P_S^B, set to 1, is searched by scanning an area in the neighborhood of the pixel from right to left, from top to bottom. The search area is a parallelogram located below P_{Left} and on its left side (the top-right vertex is on this point); the parallelogram has a width of 5 pixels and a height of 20 pixels (it is shown schematically in Figure 1.7 where P_{Left} is red colored).

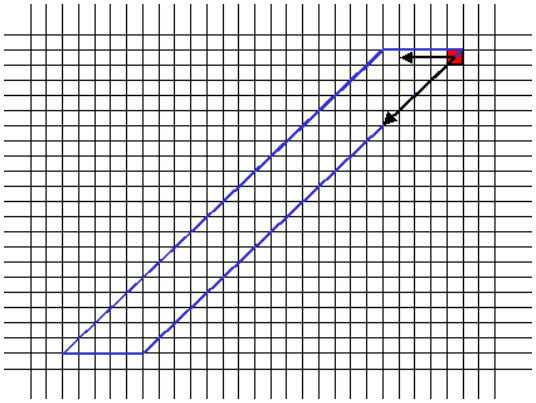

Figure 1.7: search area for the left descending path. P_{Left} is red colored

If no point P_S^B is found the algorithm reaches the end without creating any contour. Note that this is not a problem because the contour can always be recovered by integrating the paths detected by the algorithm on the other two images. As soon as a proper starting point, $P_S^B = P_S^B(i,j)$, is found, a recursive procedure begins; it considers the three $8-neighbors$ in the row below P_S^B (they are shown in Figure 1.8: they are in the positions $P_L = (i-1, j+1)$, $P_C = (i, j+1)$ and $P_R = (i+1, j+1)$). If the pixel P_L is set to 1, it is recorded as belonging to the descending path and the procedure is recursively launched using it as the starting point. The same happens for P_C. P_R is recorded as pixel in a path, and the algorithm is launched from that point, only if it is set to 1 and also the two pixels on the left and on the top in its $4-neighborhood$ (they are P_C itself and the pixel P_{Rcheck} in position $(i+1, j)$) are set to 1. This is a more restrictive condition if compared with the one used for P_L and P_C: the reason is that the direction going towards the right could lead to a path inside the lung itself (remember that the procedure described here is searching for the lateral edge of the left lung). The algorithm stops if none of the conditions for all the three pixels is verified, since no path can be started from them, or if it has reached the bottom margin of the image.

Once all the recursive calls return, the result is a set $DSet_B(P_S^B)$ of pixels which have been recorded as belonging to a possible descending path starting from P_S^B. To check if $DSet_B(P_S^B)$ is an acceptable set, the algorithm looks at the location of the bottommost points in it. $DSet_B(P_S^B)$ is rejected

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & P_S^B & P_{Rcheck} \\ P_L & P_C & P_R \end{bmatrix} \qquad \begin{bmatrix} Q_L & Q_R & Q_C \\ 0 & Q_S^U & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Figure 1.8: Neighbor pixels checked during the Descending path (Left) and the Ascending Path (Right)

if they are located above the bottommost pixel in the left side of the *First edge image*, i.e. if their $Y - coordinate$ is less than the one of this last point; in this case the algorithm is rerun using as new starting point the next one found in the search area (Figure 1.7) of the parallelogram described before and scanned in the same directions. If all the sets created after restarting the algorithm from all the possible starting points are discarded, the overall algorithm reaches the end without creating any contour.
When a set is finally accepted, there is the need to select from its pixels the proper ones to be used to create a unique contour. To this purpose the same recursive algorithm from the bottom to the top is run, using as starting point, Q_S^U; it is the rightmost pixel among the bottommost in the $DSet_B(P_S^B)$. Once the "ascending" set $ASet_B(Q_S^U)$ is created, the intersection of the two sets obtained is considered; a unique chain of pixels that constitutes the lateral contour is then formed by taking for each row the rightmost pixel of the intersection.
in Figure 1.9 some of the created paths are shown. On the left of the top row one example of discarded path can be seen. In these images also the path running to the top of the lungs is visible; it is obtained in the way described in the following paragraph.

Construction of the path running to the Top

The procedure used to detect the *Tpath* is aimed at the detection of a continuous contour running from the point $P_S^T = P_S^B$ to the top of the lung. A similar recursive algorithm, just in ascending order, is applied to the input images; as before its output is a set of points $ASet_T(P_S^T)$. For each point considered $X = X(i, j)$, and already included in the set, the algorithm now looks at the two $4 - neighbors$ located above it (i.e. in the position $(i, j - 1)$) and on its right side (in position $(i + 1, j)$). These directions are chosen according to the shape of the borders of the top of the lung. Both these neighbors are inserted in the $ASet_T(P_S^T)$ if they are set to 1, and the algorithm is recursively launched using them as starting points. The algorithm stops when the top margin of the image is reached or both the neighbors are set to zero. Once all the recursive calls return the topmost points in the $ASet_T(P_S^T)$ are checked. An error is detected if they are located not further then 20 lines above P_S^T, since in this case

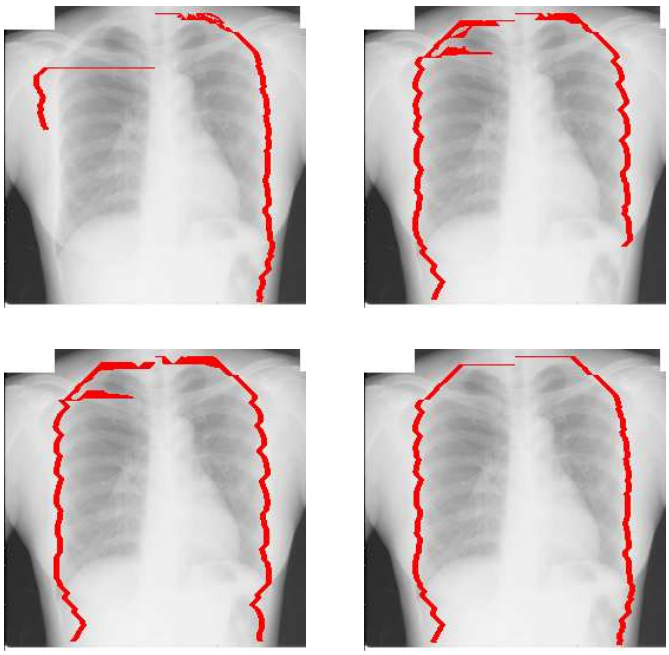

Figure 1.9: Contours created on the binary images L_σ ($\sigma = 1, 2, 3$) - Final contour (bottom right)

the $Tpath$ it's too short and hence it is probably not reaching the top of the lung field. This happens mainly when the input image does not have continuous vertical edges due to the presence of the clavicles, whose edges are superimposed to the lung borders: they show up as horizontal edges that interrupt the vertical ones. To overcome this problem, the path just created is stored, and the ascending algorithm is restarted from a new starting point. It is found by scanning (from left to right and from the bottom towards the top of the image) a search area located above the topmost point reached; this area is a parallelogram similar to the one used to find P_S^B but whose dimensions are the half (the search area and its location with respect to the topmost point are shown in Figure 1.10, where the last point reached is red colored). If no point set to 1 is found the algorithm is stopped definitely. Otherwise, the new point is connected with a segment to the topmost point reached and the algorithm is restarted; the output of this run is a set that will be added to $ASet_T(P_S^T)$.
To create the $Tpath$ from the set $ASet_T(P_S^T)$ obtained, the rightmost among the pixels belonging to it is selected for each row. The topmost pixel in the $Tpath$ delineated is often above the point at the top of the lung (this is generally called the *Apex point* of the lung). Since the apex point is generally located where the border of the lung is horizontal, to find it the absolute value of difference of the $X - Coordinates$ of points located in consecutive lines is computed. The location of the top point (t_σ, where σ indicates the

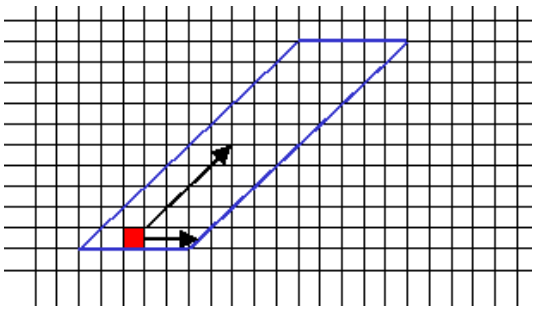

Figure 1.10: search area for the left path going to the top. The topmost point reached is red colored

path calculated on one of the three input images) is then found when this difference is bigger than a certain value, that has been experimentally chosen and differs according to the input image: it is proportionally dependent on the scale used to create the input image L_σ (equal to 14, 21, 28 for L_1, L_2, L_3 respectively).

A final contour running from the top till the bottom is created by connecting the *Bpath* and *Tpath*.
The contours created on each image L_σ ($\sigma = 1, 2, 3$) are shown in Figure 1.9. The fact that one of the three paths may be interrupted is not a problem since at least one of the other two is all defined (this happened for all the 409 images on which the algorithm was run). The procedure used to compose them can build a final contour just using the available paths.

Combining the paths detected

Since the top points $t_1 = (x_{t_1}, y_{t_1})$, $t_2 = (x_{t_2}, y_{t_2})$, $t_3 = (x_{t_3}, y_{t_3})$ of each path may not have the same vertical position, the vertical coordinates of the topmost point, Top, in the final border is set to be the mean of y_{t_1}, y_{t_2}, y_{t_3}. The contour is then created by selecting for each row, below the one at the vertical coordinate Top, the leftmost among all the pixels belonging to the different paths available. In the following the obtained contour will be referred as the *Left border path*. Repeating the same operations for the right side a *Right border path* is easily created. In Figure 1.9 (bottom right) the *Left border path* and *Right border path* obtained are visible; they are created by combining the contours showed in the other three images.

1.3.4 Edge refinement and final contour delineation

This section describes the method used to integrate the information present in the *First edge image* and in the *border paths* just presented, to get a final definition of the lung field. This is necessary since both the techniques

employed are not able to define precise contours. The *First edge image* contains borders that are not continuous; moreover some contours, such as the edges of the costophrenic angles, are often missing; another problem is that it may still contain edge pixels that are not belonging to the borders of the lung field. The advantage of this image is that those edge pixels corresponding to the lung borders define a very precise contour. On the other hand the contour obtained with the edge tracking algorithm, even though less precise, is always continuous and running from a vertical position, that is a good approximation of the top of the lung field, to the bottom of the image. Observe that the points characterizing the costophrenic angles are still uncertain since the contour, built by the edge tracking algorithm often runs till the bottom margin of the image, meanwhile the bottommost points in the left side of the *First edge image* may not correspond to the costophrenic angle. Again only the operations on the left lung will be described.

Before applying any integration procedure, the edges in the *First edge image* still need a refinement. The first thing to do is to add back to it the regions, saved in the *Little regions image*, which are located on the left side of the axis and below the topmost point, Top, of the *Left border path*.
Moreover, all the pixels in the *First edge image* located above Top are deleted. This is because the *Left border path* is always a good upper bound of the lung top.
The next step is based on the observation that the initialization point, P_{Left}, has been detected as the point of the left lung nearest to a line oriented at 135° and passing by the origin. All the points above the line with the same orientation and passing through P_{Left} can then be discarded from the *First edge image*, since they surely do not belong to the lung field. In this way the procedure is able to discard the edges belonging to external bony structures, such as clavicles, that may still be present in this image.

A constraint for the possible vertical position of the edges of the costophrenic angle is defined by searching the vertical coordinate of the bottommost point in the intersection of the *First edge image* and the *Left border path*. The deletion of the edge pixels below this vertical coordinate in the *First edge image* is the last step before proceeding to the integration of the edges it contains and the *Left border path*.

The integration is carried out by following the borders in the *First edge image*, starting from their topmost point and going in a descending direction, to search for a continuous path; when a hole is found, it is filled by taking the pixels in the *Left border path*, whose advantage is to be always continuous. The procedure stops when it reaches a point L_{angle} located on the same row of the bottommost points in the *First edge image*, since it is obvious that the rest of the *Left lateral contour* is the one belonging to the *Left border path*. As required during the creation of the path running to

the bottom (see paragraph 1.3.3) the algorithm is run also in a bottom-up direction, and starting from the point L_{angle} itself, to get a more precise definition of the contour.

On the right side of the left lung, nearby the spine, the aim is to have a rough outline of the dorsal column. To do this a line parallel to the axis is created, which passes by a point, in the *First edge image*, that is located on the right border of the lung: this is the point that is the nearest to the axis. This last point is then joint with a segment to the topmost pixel in the *First edge image* in order to create a closed contour on the right side of the left lung (this segment, together with the line parallel to the axis, are drawn with the color blue in the images in the top rows of Figure 1.11).

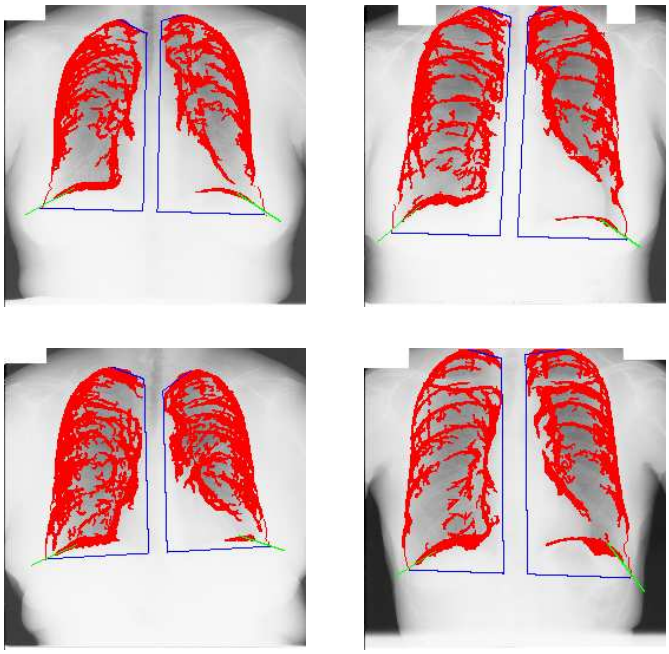

Figure 1.11: Segment used to define the costophrenic angle

1.3.5 Costophrenic angles detection

To complete the contour, the positions of the costophrenic angles must be defined. They are needed to create a bottom boundary for the area of the lung field; this is indeed defined by the segment connecting the points corresponding to the two costophrenic angles. These points are obviously located at the junction of the lateral borders of the lungs and the borders of their bottom (these last borders are defined by the presence of the diaphragm). The edges of the bottom of the left lung are selected from the *First edge image*; they are found by searching the region on the right side of the pixel

in the position of L_{angle}, whose bottommost point is the nearest to L_{angle} itself. The 5 leftmost and bottommost points of this region, are fitted with a polynomial fitting method that minimizes the $\chi - square$ error statistic. The purpose is to find a line l interpolating them. The intercept of l is then modified in order to create a parallel line, passing through the point with the highest $Y - Coordinate$ among the five selected (see the images in the top row of Figure 1.11, where l is drawn with the green color). The intersection between this line and the *Lateral contour* detected defines the point L^1_{angle} where the costophrenic angle is located: the contour is ended in L^1_{angle}.
Repeating the same procedure for the right side the point R^1_{angle} is found. On this side the edges are often weaker, due to the presence of the heart. This could cause problems in detecting the bottom part and hence the proper position of the right costophrenic angle. Since our algorithm is always very precise in detecting the point L^1_{angle} this information is employed to check the position of R^1_{angle}. The underlying assumption is that L^1_{angle} and R^1_{angle} should be approximately located in a symmetric position with respect to the axis. This is the reason why the check on R^1_{angle} is performed by computing the absolute value of the difference between its vertical coordinate and the one of the point, L^R_{sim}, that is symmetric to L^1_{angle} with respect to the axis (and is of course located on the right side). If this distance is bigger than 15 pixels and L^R_{sim} is located below R^1_{angle}, we substitute to R^1_{angle} the point on the *Right lateral contour* that is in the same row of L^R_{sim}; in this way the continuous lateral contour is stopped also in the right side.
To close the contour a segment is defined that connects the two costophrenic points detected in the two lungs.

1.4 Results

1.4.1 Results on the JSRT database

To evaluate the quality of the contour obtained we considered as error a part of the lung that is not properly enclosed into the segmented lung area or a significant part, not belonging to the lung, but included into this area. All the segmented images in the database have been divided into three classes, *Perfect*, *Little Errors* and *Bad Errors*, according to the errors they contained. Images classified as *Perfect* are the ones where the lung field perfectly encloses the lungs, adhering to their real borders; some examples are shown in Figure 1.12. The images in the *Little Errors* class are 8. Some errors are due to small parts of the lung that are not included; some others have been detected at the bottom of the right lung, where the segmented area includes a part, not belonging to the lung field, which can be rich of edges due to organic gasses; they are shown in Figure 1.13. Images contain-

ing *Bad Errors* are those with big parts of the lung not included into the segmented lung field or they are due to extra parts, very big, not belonging to the lungs and included into the segmented lung field. No bad errors have been detected in the images of the *JSRT* database.
In the table 1.1 the results described are summarized and their percentage is given. To be more specific they have been measured separately on the left and right lung (first and second row); the third row gives an overview on both the sides of the lung.
We compared our method with the only one developed and tested on the same database. As already said this method, described in [85], does not include in the lung area the parts behind the diaphragm, the spinal column and the heart. The same qualitative criteria were used to judge the results it produces, and 50 errors have been detected. Almost all of them are due to missing parts; precisely we have classified 10 images as belonging to the *Bad Errors* and 40 images as belonging to the *Little Errors* class. Moreover it can be said that the contour detected is not as precise as the one created by our method since it is often an internal contour. Note that this method is the one, presented so far in the literature, with the best performance.

Examined Part	**% Perfect**	**% Little errors**	**% Bad Errors**
Left Lung	99.6	0.4	0
Right Lung	97.1	2.9	0
Both Lungs	96.7	3.3	0

Table 1.1: Results on the *JSRT* Database (247 radiographs)

1.4.2 Results on the Niguarda Database

Our method has also been tested on the 162 images of the *Niguarda* database and the results are summarized in table 1.2; 5 images are in the *Bad Errors* class; they are shown in Figure 1.14. The images in the *Little Errors* class are 8 and are shown in Figure 1.15. Examples of the Perfect results obtained are shown in the Figure 1.16.
Note that no method already presented in the literature has been tested on such an amount of images belonging to different databases.

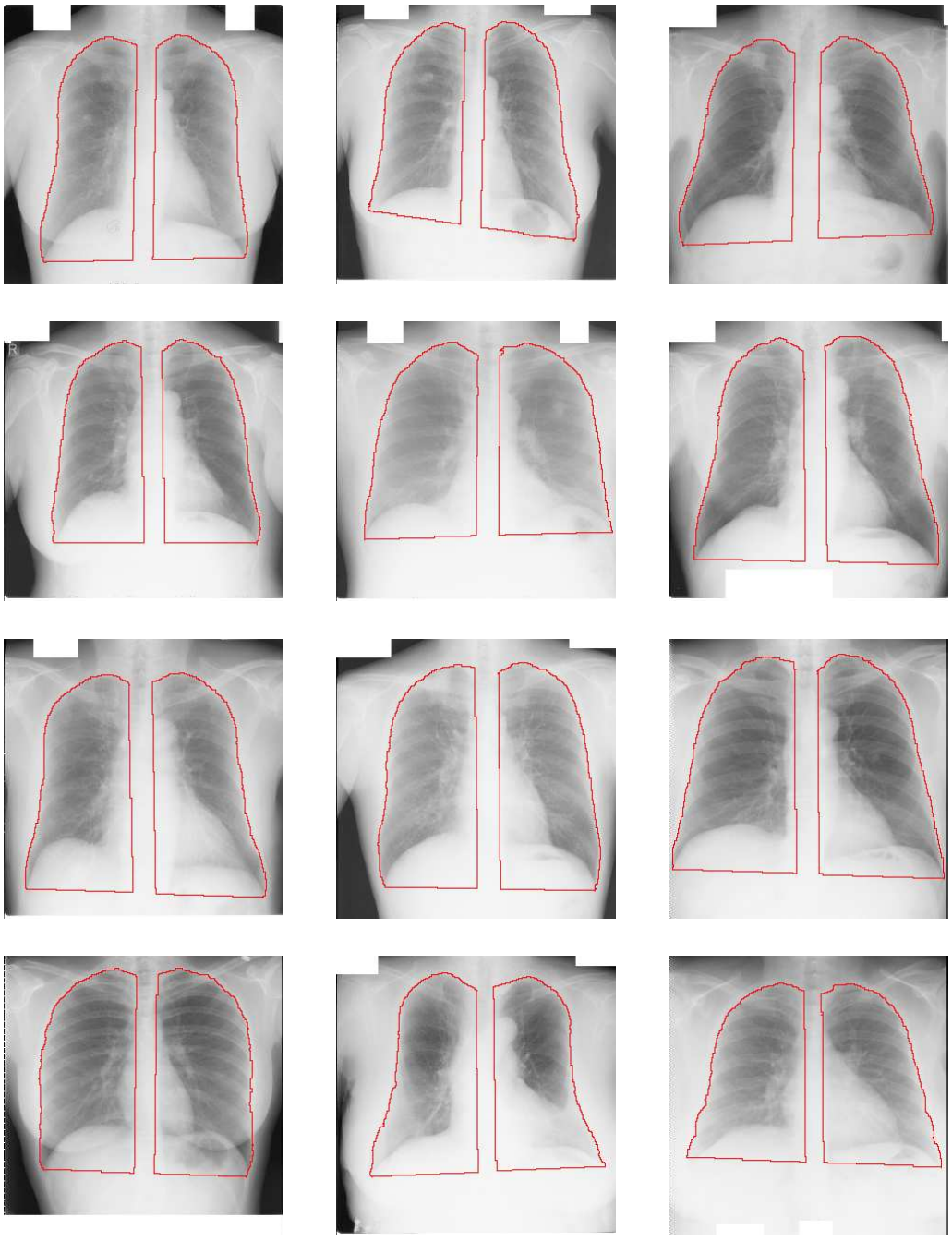

Figure 1.12: Examples of the perfect results obtained on the *JSRT* Database

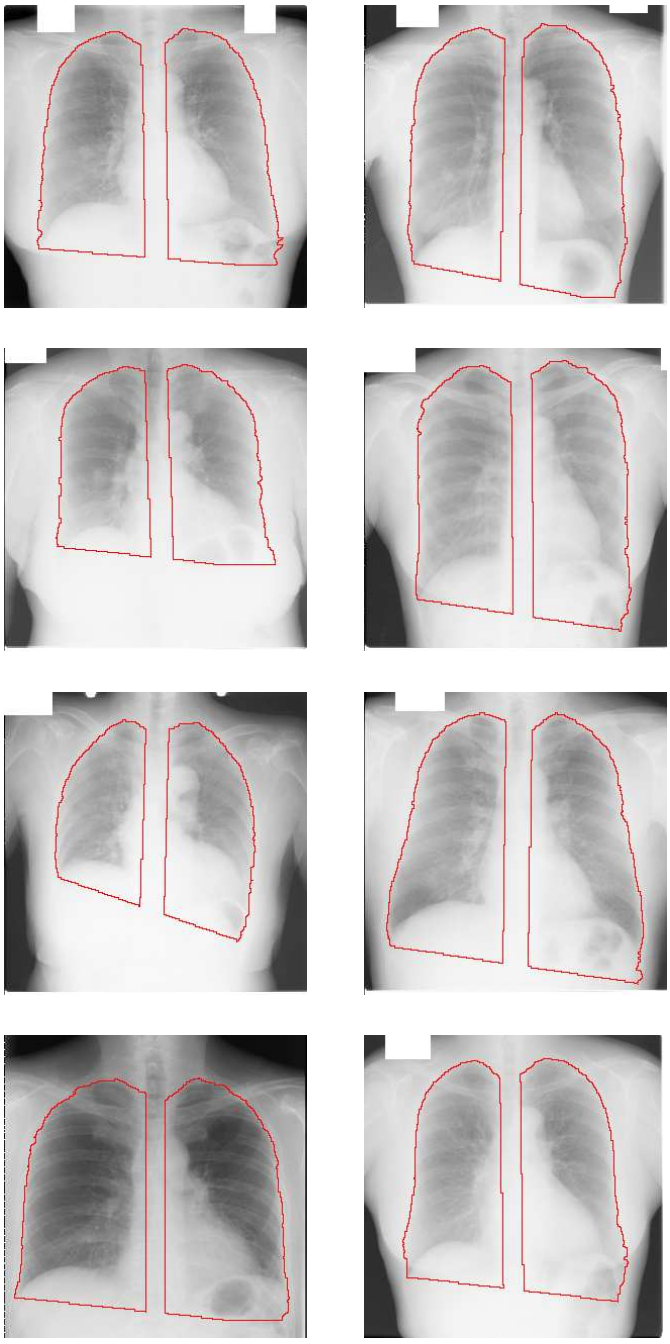

Figure 1.13: The 8 images containing Little errors in the $JSRT$ Database

Examined Part	% Perfect	% Little errors	% Bad Errors
Left Lung	96.3	1.2	2.5
Right Lung	93.8	4.9	1.3
Both Lungs	92.0	5.0	3.0

Table 1.2: Results on the *Niguarda* Database (162 radiographs)

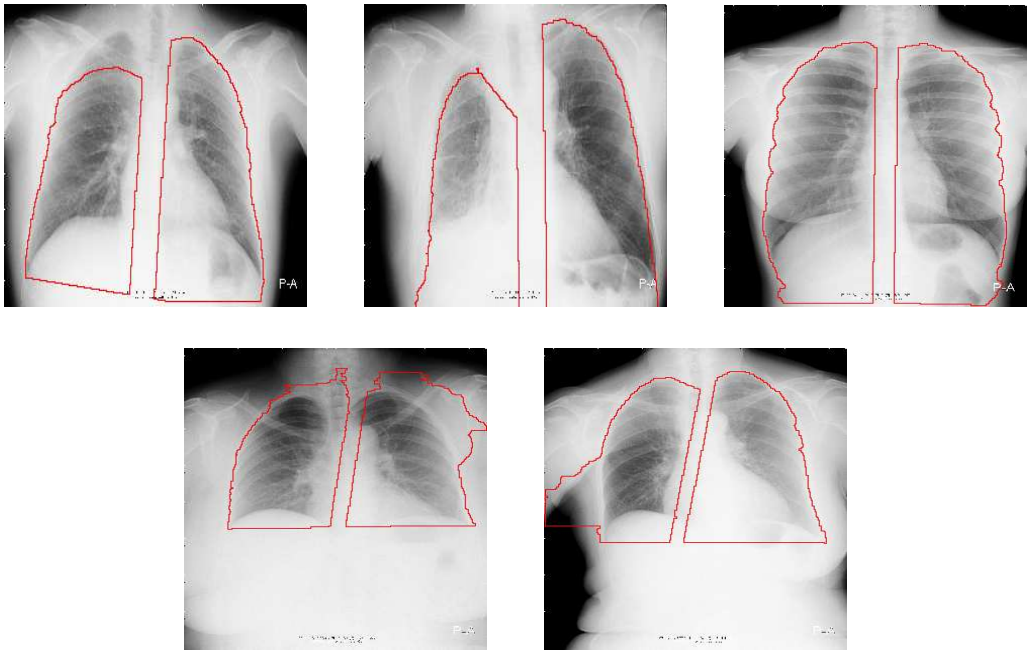

Figure 1.14: The 5 images containing Bad errors in the *Niguarda* Database

1.5 Segmentation of the visible lung area

1.5.1 Purpose

The aim of this section is to describe a further step which is added to the previously described segmentation algorithm, to give a better definition of the lung field, hence allowing to retrieve more information about the lung structure itself. Its purpose is to separate, for each lung, the *visible parts* from the ones hidden behind other structures such as the heart, the diaphragm and the spine. The *hidden areas* are completely different from the *visible areas*: they are characterized by much brighter grey levels and a more uniform texture. A differentiated analysis of the two areas may be helpful for any kind of computerized analysis of the lungs. The separation between them has indeed proved to be efficient to get a better characterization of the *candidate regions* detected by our computer aided system for the detection of lung nodules (see Chapter 3).

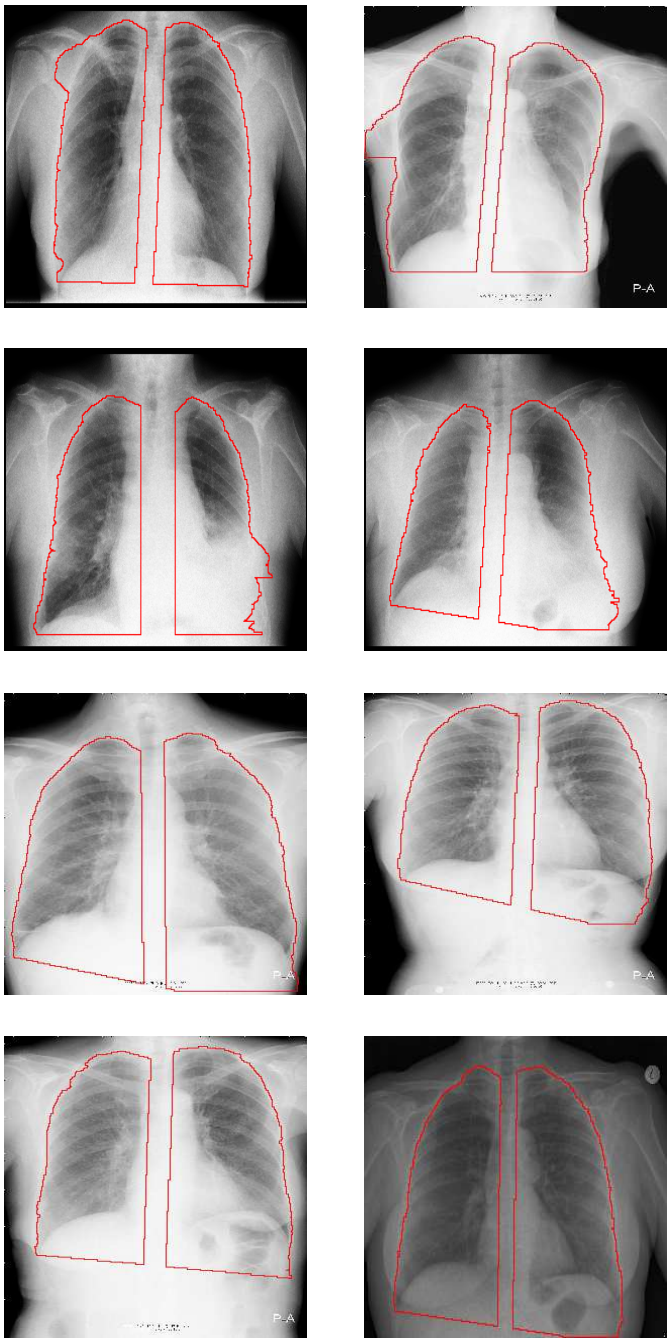

Figure 1.15: The 8 images containing Little Errors in the *Niguarda* Database

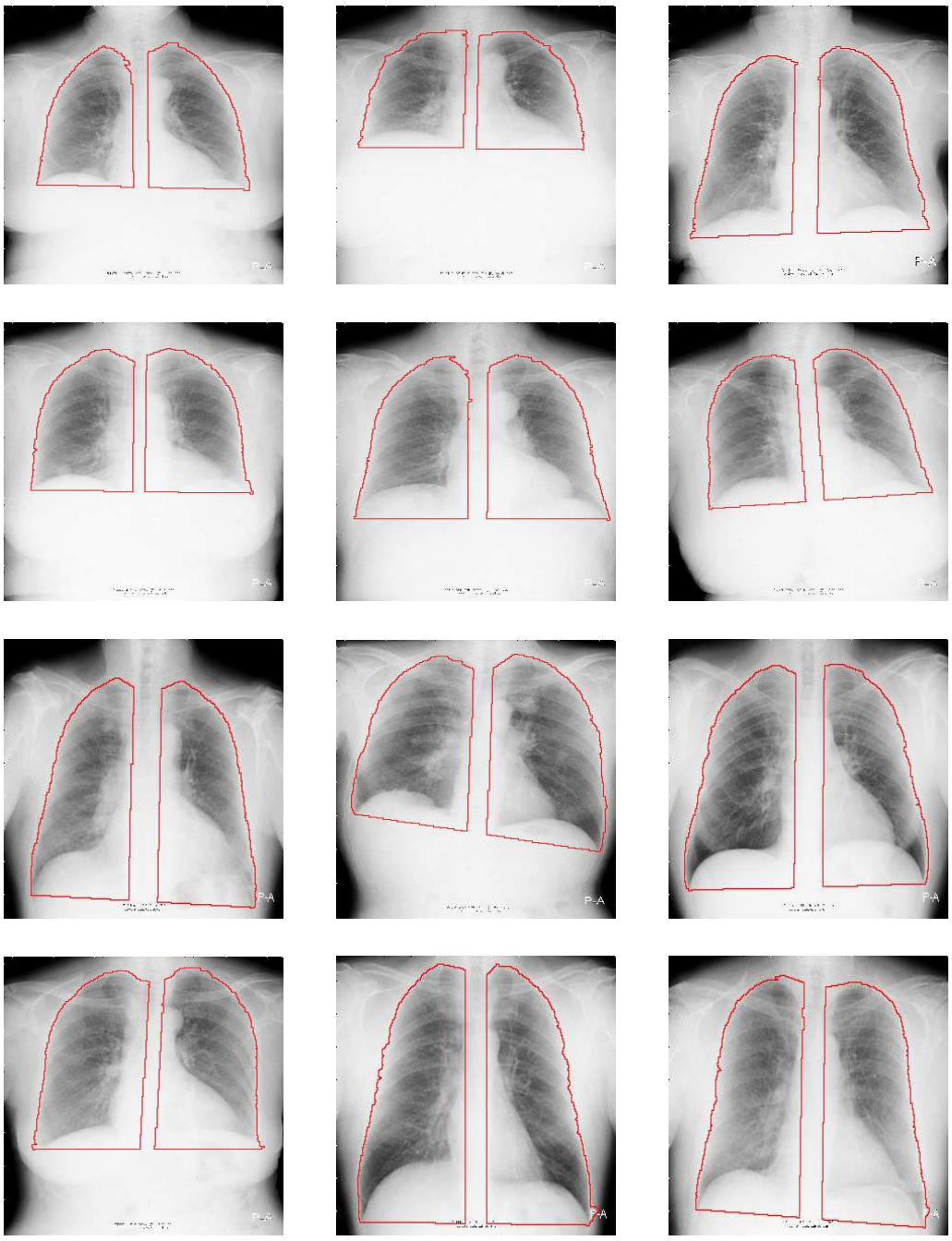

Figure 1.16: Examples of the perfect results obtained on the *Niguarda* Database

Moreover the analysis of the *hidden areas* has been helpful to detect and correct the errors, made by our previously described segmentation algorithm, that are caused by external structures, such as gases, in the right side of the radiographs. Since the two areas to be separated differ mainly for the grey levels of their pixels, the method starts with a clustering algorithm, described by Arbib and Uchiyama in [82] and based on competitive learning. Each pixel is assigned to a certain cluster according to a distance measure based on the grey level of the pixel itself. The result is a first rough separation between the two areas. The segmentation is improved trough the use of simple derivative filters.
The input of the algorithm is a preprocessed version of the original radiograph, at its original dimension and with 4096 grey levels; the preprocessing has the purpose of removing some noise. To this aim a median filter of 5 pixel is applied, followed by the convolution with a gaussian filter with standard deviation $\sigma = 3$. The grey levels are then scaled to the range $[0, 255]$ and the image is down-sampled to the dimension of 512 by 512 pixels. In the following steps this preprocessed image will referred as the *Original Image*.

1.5.2 Clustering method

The result of the segmentation algorithm, described in the previous paragraphs, allows to apply an enhancement procedure, and all the steps needed to perform this further segmentation, just to the lung tissue and separately on the left and right lung. An *enhanced image*, where the borders of the *visible lung areas* appear more visible is thus created by applying an histogram equalization (see [67]) separately on the left and right lung in the *Original image*. This operation is followed by a simple *Min-Max* filter to enhance the image contrast by means of a non linear, extreme value sharpening technique, whose effect is to increase the contrast where the boundaries between objects are characterized by gradual changes in the grey levels. Its definition is the following:

$$GN(x,y) = \begin{cases} \max & \text{iff} \quad |\max - G(x,y)| \leq |\min - G(x,y)| \\ \\ \min & \text{otherwise} \end{cases} \tag{1.1}$$

where min and max are the minimum and maximum grey values computed on a window $\mathrm{Win}(x,y)$ centered in (x,y). The window size used is 3 pixels. The resulting image is the input of the clustering algorithm to search for three clusters. This decision is due to the fact that the visible area is characterized by a darker mean grey level, the part behind the heart is often characterized by lighter grey levels, similar also to the ones of the pixels in the regions of the gases, and finally the part behind the diaphragm contains pixels with the highest grey levels. In Figure 1.17 some examples of the results of the clustering method are shown on the left side; all the pixels

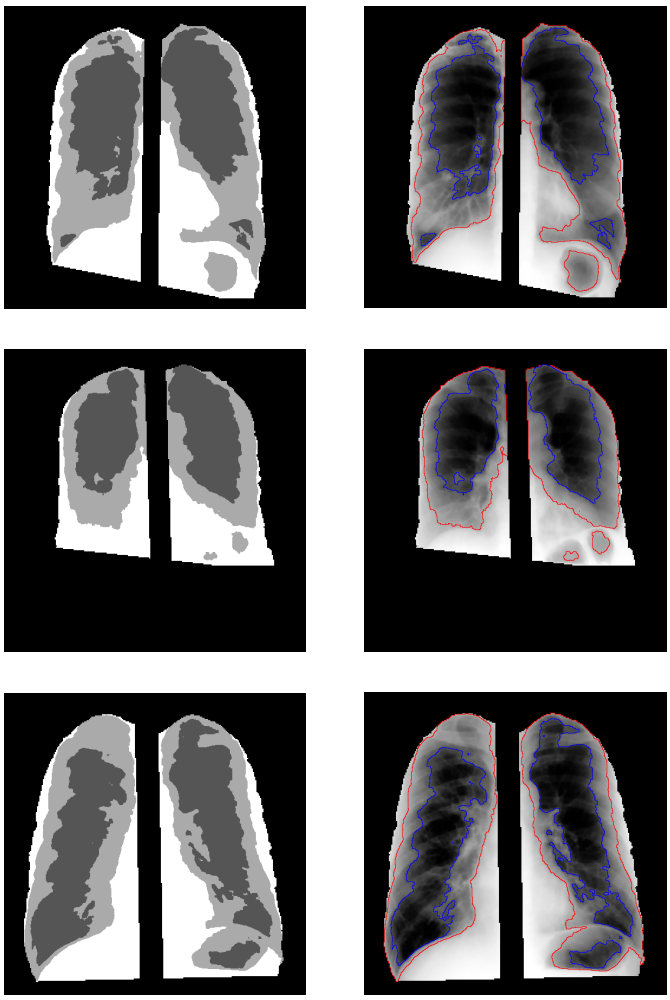

Figure 1.17: Clustering results

in the same cluster have the same value $CValue \in [1, 2, 3]$ ($CValue = 1$ corresponds to the cluster whose representative has the lowest grey value with respect to the other clusters, $CValue = 3$ corresponds to the cluster represented by the highest grey value). In the right column the borders of the three clusters are colored, with red and blue color, on the *enhanced image*. For each lung the biggest region formed by the union of the two clusters with $CValue = 1$ and $CValue = 2$ is selected; filling it so that it touches the external borders of the lung, a first *Mask of the visible areas* is obtained. As shown in the top row of Figure 1.18, for some of the images the clustering method well defines the *visible areas*, for other images it is not enough, especially at the bottom part of the right lung. Two examples are shown in the bottom row of the same Figure; they are typically caused by the missing costophrenic angles or by the fact that some external structure, such as internal gases or the spinal column, may still be included in the *Mask of the visible lung area*; this last must then be refined with the operation explained in the next section.

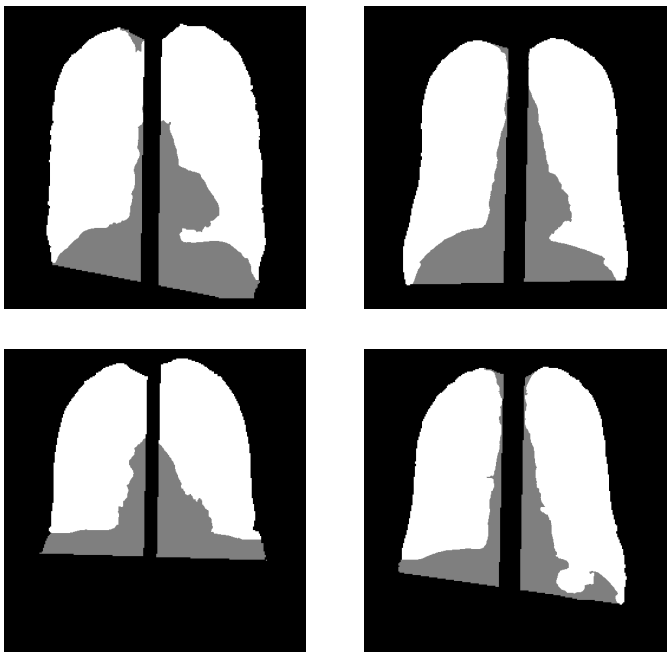

Figure 1.18: The mask obtained selecting using 2 Clusters in the result of the clustering algorithm

1.5.3 Edges detected with derivative filters

The *Mask of the visible lung areas* often contains errors that are mainly located at the bottom part or in the part nearby the spinal column. To correct them the results of simple derivative filters have been used. The aim

is to search for the most visible edges at the bottom and nearby the spine; they will be used to discard the extra parts still contained in the *Mask of the visible lung areas* or to add the parts lost. The horizontal edges, such as those between the visible part of the lung and the diaphragm, are obtained by filtering the *enhanced image* with a vertical Sobel filter, $YSob$; the mask used for this filter is shown in figure 1.19. The vertical edges that separate the *visible lung area* from the part behind the spinal column, are obtained filtering the *enhanced image*, separately on the left and right lung, with two horizontal Sobel filters; in Figure 1.19 the one used for the left lung, $XSobL$, is shown; the right lung is filtered with a filter whose mask is $= -XSobL$. To get two binary images, containing the horizontal edge regions and the vertical edge regions respectively, the procedure selects (separately on the two lungs) the 5% of the pixels with the highest value of the corresponding derivative image. Deleting from the resulting binary images all the connected regions with less then 100 pixels the *horizontal edge image* and the *vertical edge image* are formed. Finally a *total edge image* is obtained by keeping all the connected regions in the *horizontal edge image* and selecting from the *vertical edge image* those regions that intersect at least one of the regions in the *horizontal edge image*. The edges thus selected always contain the borders of the bottom part of the *visible lung area* and some of the vertical edges attached to it. Some unwanted edges belonging to the intestinal gases or other structures may still be present. They can be easily recognized for their shape and location with respect to the *Mask of the visible areas*; based on this fact, the decision is to further discard the connected regions in the *total edge image* which do not intersect the *Mask*, those regions with a convex shape, and the ones with a rectangularity factor bigger then 0.33. Some examples of the *total edge image* before and after this cleaning are shown in Figure 1.20; the deleted edges are signed with a red arrow in the figures in the left column.

$$YSob = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \qquad XSobL = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

Figure 1.19: Sobel filter used

1.5.4 Integration of the two techniques

The edges in the *total edge image* are employed to refine the *Mask of the visible areas*. As before the process is applied separately on the right and left lung. In the following it is described just for the left, since on the right lung the operations are the same but opportunely reversed.
On the bottom part, the bottommost region (*bottom edge region*) is ex-

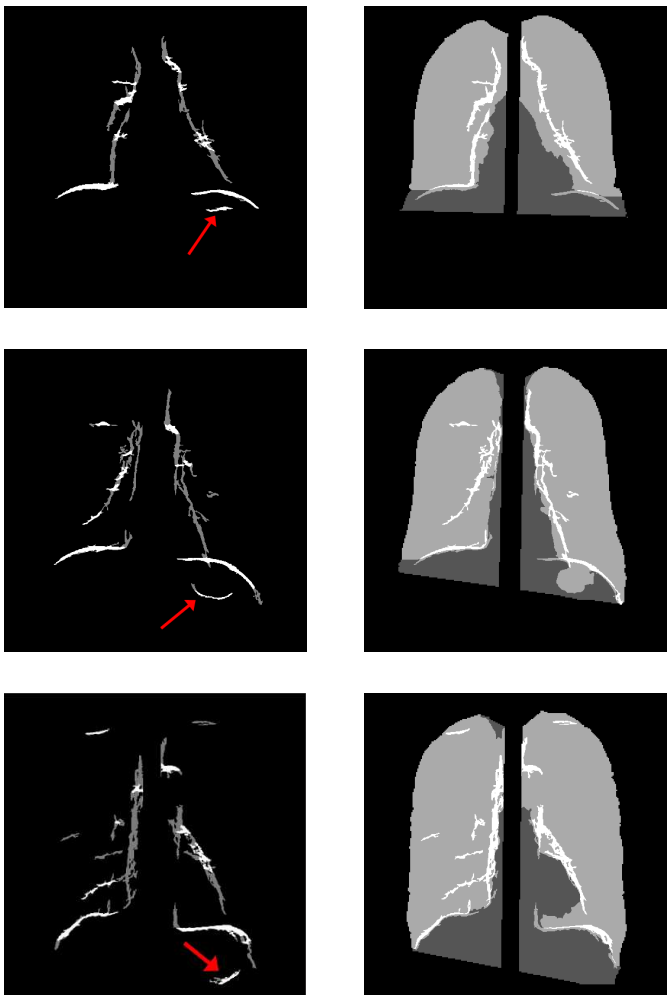

Figure 1.20: Borders detected

tracted from the *total edge image*; it is the one corresponding to the bottom contour of the *visible lung areas*, and defined by the upper border of the diaphragm. Deleting all the pixels in the *Mask of the visible areas* below it, all the extra parts in the bottom are deleted; in the same time the pixels above it are added to the *Mask*, if they are not already present, to repair the fact that some parts could be missing.
The bottom edge region is used also to estimate a new point corresponding to the position of the costophrenic angle. This is done by finding the point where this region intersects the external border of the lung field; if this doesn't happen the new costophrenic angle position is estimated by selecting 1/4 of the points of the skeleton of the bottom region which are the most external (they are the leftmost for the left lung and the rightmost for the right lung) and then fitting them with a polynomial fitting procedure that minimizes the $\chi - square$ error statistic. The purpose is to find a line l interpolating these points. The intercept of this line is modified to create a line, parallel to l, that passes through the bottommost point of the bottom edge region. The intersection of this parallel line with the external border of the lung field determines the new costophrenic angle. The *Mask of the visible areas* is then filled above this line to have a close region. In the left column of Figure 1.21 some examples are shown. Note that this simple procedure may produce some jumps in the contour of the bottom area (see the right lung in the second row). To overcome this problem and refine also the borders of the visible lung area nearby the spine, a very simple contour following procedure is used; this will be explained in the next section.

1.5.5 Contour following to get the final segmentation

To end the refinement procedure and get a good separation between the two parts of the lung, a final and very simple contour following procedure is used. The edge regions in the *total edge image* are often not so well defined since the borders nearby the spine are characterized by a smooth and gradual change of their grey levels; this is because, in this part of the thorax, the lung and ribs "turn" gradually to become the anterior part of the lung and ribs. For these reason the contour followed are the ones obtained by applying to the *Original Image*, input of this algorithm, a locally adaptive scaling operator (described in [43]) whose purpose is to produce an enhancement of low contrast edges, especially inside the bright image domain. The decision to use this operator is indeed due to the fact that this area is characterized by pixels with the highest grey values. This is a local operator working on squared windows whose size has been set to be of 5 pixels. Its output has been histogram equalized, separately on the two *visible lung areas* already defined, to get a better enhancement of the edges. The final result, i.e. the *locally enhanced* image, is shown in the right column of Figure 1.21. Note that no histogram equalization has been applied to the hidden lung areas

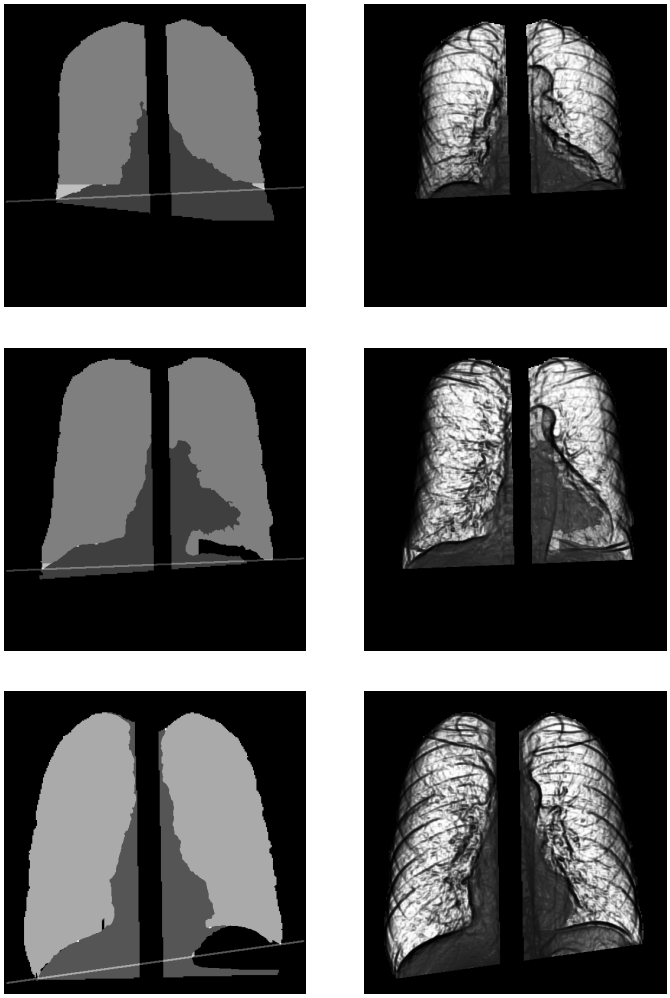

Figure 1.21: *mask of the visible lung area* before the final contour following (Left) - *locally enhanced image* (Right)

since the purpose of this step is to discard from the *Mask of the visible lung areas* those regions of the hidden area that are still included; what is needed then, is to detect and follow the edges of the hidden areas still contained in the mask. Again the edge selection and the contour following are described for the left lung, since the operations are the same, but reversed, for the right lung.
A *binary image* containing these edges is formed by selecting the 50% of the pixels in the *Mask of the visible lung areas* with the lowest grey value. Starting from the costophrenic angle, the contour of the *Mask of the visible lung area* is followed from left to right to define a path without jumps. A jump in the contour is found when the absolute value of the difference between the $Y - coordinates$ of the pixels in the contour, and located in consecutive columns is bigger than 5. A jump is repaired by taking, from the *binary image*, the nearest pixel to the previous in the contour, and obviously located in the column of the jump. The procedure goes on from this last pixel and stops when the vertical position Y_{Stop} is reached: this is the vertical coordinate of the point in the *visible lung area* which is the nearest to the axis. In the Left column of Figure 1.22 we show some of the edges in the *binary image* that are nearby the spine (superimposed on the masks of the *visible lung area*) and below Y_{Stop}. The masks obtained are shown in the right column of Figure 1.22. Notice that the jumps have been repaired and the contour is finally fixed.

1.6 Global Results and Conclusions

The method, just described, not only produces a good separation between the *visible lung area* and the *not visible one*, but it also repairs those errors at the bottom of the lung, where extra parts due to the intestinal gases are wrongly included in the segmented lung field. The errors that cannot be fixed are the ones due to the missing parts of the lungs; anyway the performance of the overall method are considerably increased with respect to the ones reported in section 1.4.

Regarding the images in the $JSRT$ Database, **all** the *Little Errors* have been detected and corrected. In Figure 1.23 the corrected errors in the $JSRT$ Database are shown.

On the top row of Figure 1.24 we show the 3 corrected errors in the images of the $Niguarda$ Database, previously classified as *Bad Errors*. Note that, in the first image, just the error at the bottom of the right lung is detected and corrected; the error in the left lung is not repaired since it is due to a missing part.
Figure 1.25 shows the results of the algorithm when run on the images clas-

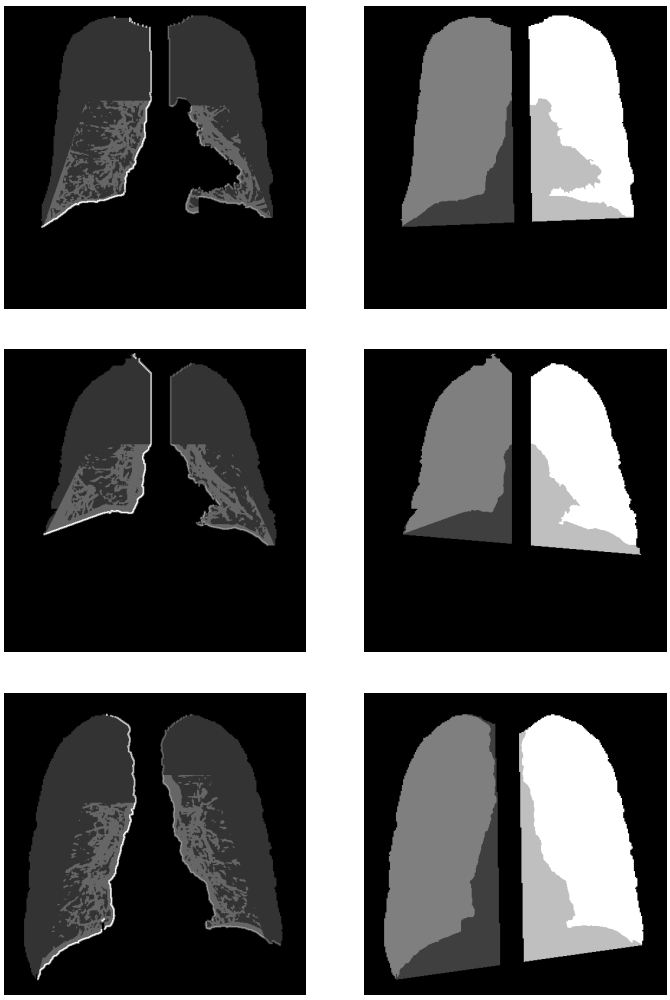

Figure 1.22: Edge used for following the contour and *Final masks*

sified as *Little Errors* in the *Niguarda* Database: 5 errors are completely corrected and 1 is partially corrected. In table 1.3 the performances of the overall algorithm on the *Niguarda* Database are summarized.

Examined Part	**% Perfect**	**% Little errors**	**% Bad Errors**
Left Lung	96.4	1.2	2.4
Right Lung	98.1	0.7	1.2
Both Lungs	95.7	1.9	2.4

Table 1.3: New Results on the *Niguarda* Database (162 radiographs)

The very good performance obtained by the overall lung segmentation method presented in this chapter, confirms its efficacy: no errors are left in the images of the *JSRT* Database and few of them are left in the *Niguarda* Database. Its robustness is also proved by the fact that it has been tested on a set of 409 images belonging to 2 different databases, containing completely different images, and the results obtained are good in both the cases; at the state of the art this is the method tested on the biggest amount of images. Furthermore it differs from several of the methods already described in the literature since it is based on no assumption such as those regarding the position and the orientation of the thorax in the radiograph, the structural abnormalities of the chest and the heart. Moreover, it not only identifies the lung area generally defined by the other methods, but also includes into the segmented lung field the bottom part of the lungs (hidden by the diaphragm) and the parts behind the spine and the heart. This fact is of utmost importance since the segmentation just described has been realized as the first step of our Computer Aided Diagnosis system described in chapter 2.

The system described has been implemented in IDL, an interpreted language; when executed on a Pentium III with 128 Mb of RAM it takes a maximum of 50 seconds. This time can be reduced significantly, to at least its 10%, by programming the algorithm in a compiled language and optimizing the code.

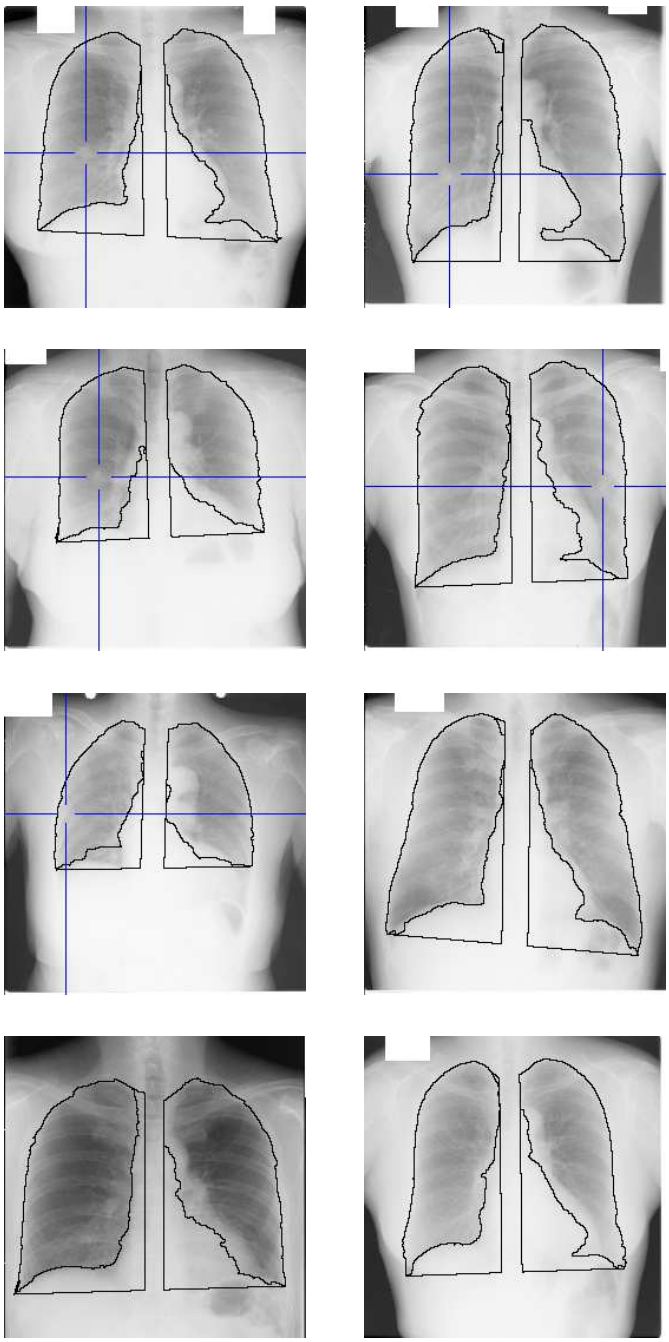

Figure 1.23: the Little errors in the images of the $JSRT$ database now all corrected

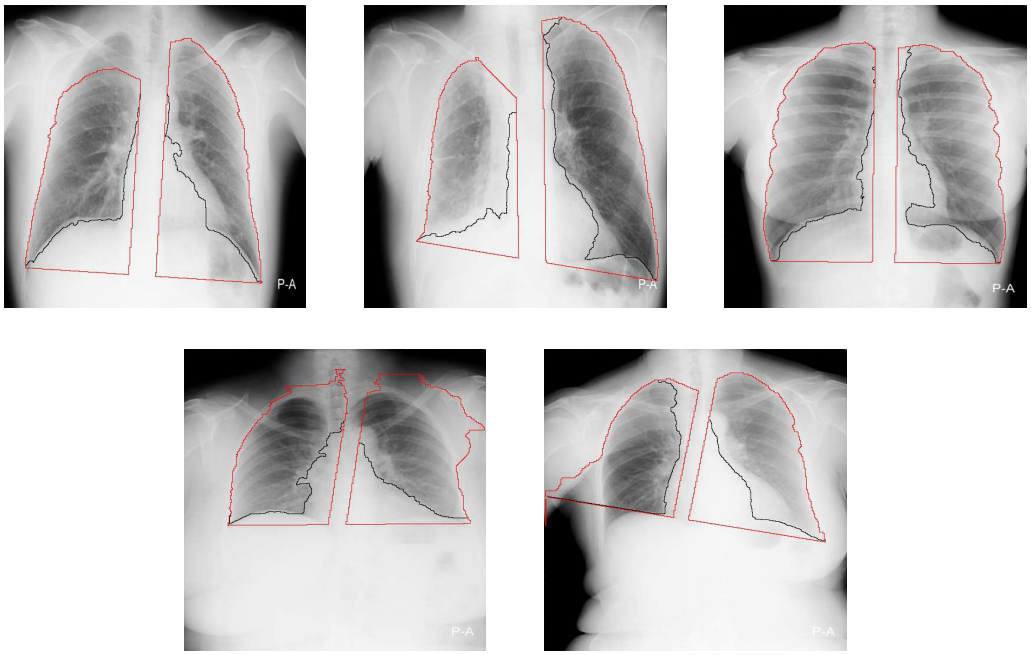

Figure 1.24: The results on the Bad Errors in the *Niguarda* Database; 3 corrected (Top row) - 2 not corrected (Bottom row)

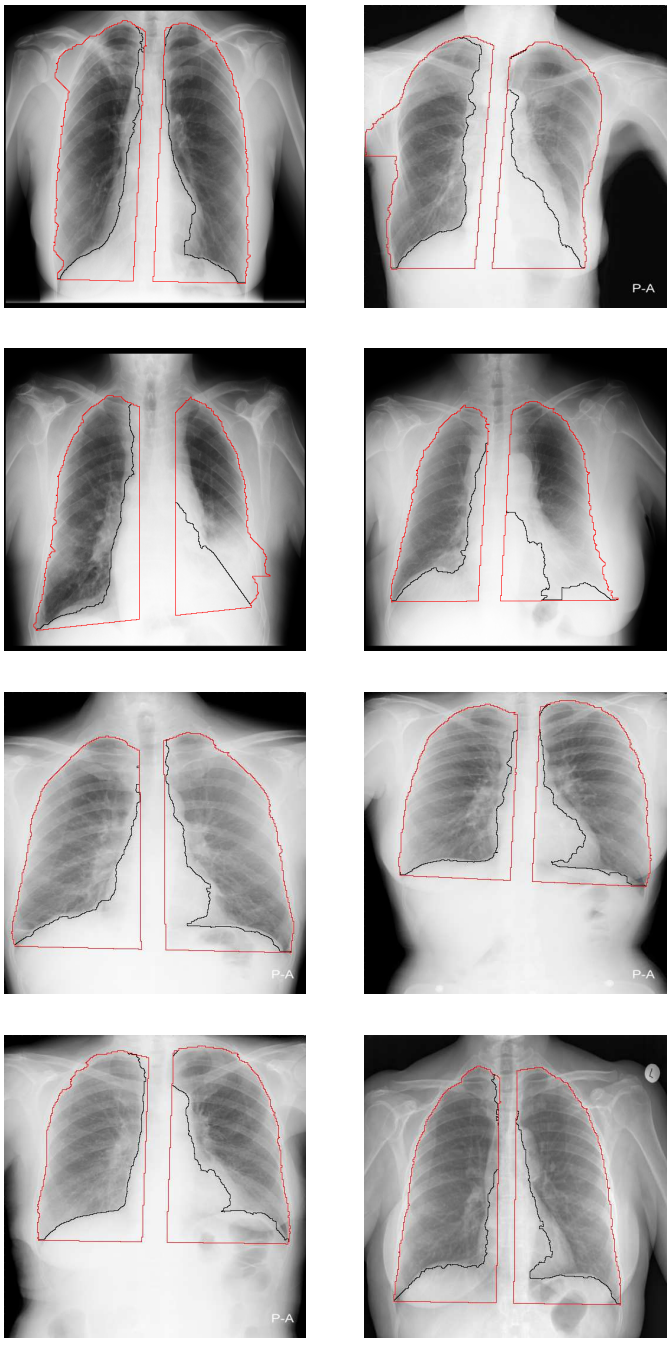

Figure 1.25: The Little Errors on the *Niguarda* Database; 2 not corrected (Top row) - 1 partially corrected (Left in the second row) - 5 completely corrected

Chapter 2

Extraction of the Nodule Candidate Regions in Postero-Anterior Chest Radiographs

Image processing techniques and Computer Aided Diagnosis (CAD) systems have proved to be effective for the improvement of radiologists' diagnosis. In this chapter we describe a system for processing Postero Anterior chest radiographs in order to detect lung tumors. It extracts a set of nodule candidate regions characterized by a high sensitivity ratio. This result is obtained by means of two consecutive multi-scale methods to enhance the visibility of the nodules and then extract a first set of candidates.
The method has been developed and experimented on the 247 radiographs in the JSRT database: 154 of patients with lung nodules and the remaining with no nodules. The set of candidates detected contains 31100 candidates, approximately 125 per image, loosing only 5 True Positives out of 154, and hence reaching a Sensitivity ratio equal to 0.97%.

2.1 Introduction

In the Introduction the reasons why CAD systems for lung tumors detection would be helpful for clinical purposes are reported and motivated.
In the last two decades a lot of research work has been focused on the development of such systems and a wide variety of them have been already proposed ([28], [27], [52], [47], [95], [38], [41], [8], [80], [63], [78], [77], [1], [100], [99], [98], [92], [39], [40], [95]). Most of the methods presented in the literature start with the extraction of a first set of candidate nodules. The aim is to extract all the regions that may contain nodules without loos-

ing too many True Positives (in the following we will call *True Positives* the candidates that are real nodules, and *False Positives* are the regions wrongly extracted as candidates). This is generally done by means of a two steps procedure which first enhances the visibility of the nodules, and then extracts the candidates.
The enhancement procedure is generally done by subtracting a background image from the original one, so that the regions characterized by the highest frequencies are detected (see [26] and [51] for some examples). Second, all the regions that may contain nodules are extracted. Since a nodule is generally considered as a bright circular region with high contrast, the candidates extraction is generally performed with methods based on rules regarding shape, grey level and contrast of the regions. Examples are the methods that apply to the nodule enhanced image a template matching procedure or a simple thresholding.
Since the focus of these first two steps is to detect all the nodules, if present, the resulting set of candidates extracted is generally quite high and must be reduced. Systems trying to extract a set with less candidates have been developed, but their high True Positive loss has lead to approaches which leave to proper classifiers the task of reduction. This problem will be discussed in the next chapter.
In this chapter an approach to enhance the visibility of the nodules, and then extract a set of candidate regions, is presented. The comparison with other two methods, presented by Ginneken and Shilham in [1] and by Keserci et *al.* in [8], that are the only ones tested on the same database, proves the very good performance of our method, both in terms of cardinality of the set extracted and sensitivity ratio (See Appendix C). The system reaches a sensitivity ratio equal to 97% and a set of candidates extracted with a cardinality equal to 31100 regions, about 125 per image.
The novelty of the method is the use of two multi-scale algorithms to first enhance the visibility of the nodules and then extract a set of candidates. Both the multi-scale schemes are based on the fact that the nodules to be detected have different sizes and are characterized by different texture, dependent on their location in the lung. As an example, nodules nearby the hilum appear with a more spiculated shape than the ones in the upper part of the lungs, or located in the parts hidden behind the diaphragm; moreover their texture is less homogeneous. Another difference can be noticed in the grey levels of the pixels of nodules in the parts behind the spine or the diaphragm; these have much higher grey levels than the ones in the part of the lung defined as *visible* in the previous chapter.
To develop a scheme that is able to enhance the visibility of tumors with so different characteristics a multi-scale scheme is of utmost importance. As cited before, the schemes presented in the literature proceed with an enhancement that is often performed with the subtraction, from the original image, of a smoothed version where the background structures are sup-

pressed; others employ a similar procedure, generally called *unsharp masking*, based on the subtraction of a smoothed version of the image from an image where the high frequencies are highlighted (see [28], [27], [98], [100]). These operations are aimed at recovering the difference of the grey levels between different nodules, due to their background, and they should highlight the regions with a not homogeneous texture. Their result is of course dependent on the size of the operator used to produce the smoothed image: it indeed highlights only a limited range of frequencies, i.e. the details (or frequencies) belonging to a limited scale (a certain range of frequencies). It is straightforward to understand that such a method cannot be used to capture all the details belonging to nodules of different sizes and grey levels. A practical example is the enhancement procedure used by the scheme presented by Giger et *al.* in [26]. The authors use a smoothing operator of a certain size to enhance the visibility of the nodules. The candidate regions are then extracted by thresholding the result, and some rules based on shape are then used to discrad some of them. Our implementation of the same method on the $JSRT$ Database showed its weakness: lots of nodules were lost, and the ones detected were just some of the tumors with the biggest size. Some other, with the same size but with spiculated texture characteristics where lost, since their grey levels frequencies belonged to the highest ones not detected by the filter. This suggested the use of a multi-scale method for both the steps of enhancement and extraction.
The framework of a general multi-scale scheme is often a function, parameterized by a scale parameter σ, which is able to extract from the image the information belonging to a limited range of frequencies (scales) strictly related to σ itself. Repeating the procedure for different values of the parameter, the multi-scale scheme is able to extract all the information belonging to the different scales. The final result is then created by the integration of all the results obtained at the various level of details considered.

2.2 Materials

The method has been developed and tested on the images in the $JSRT$ Database. The application of the method on the 162 images of the $Niguarda$ database is useless since no clinical information, regarding the presence of nodules, is available for these radiographs.
Before processing, the images have been down-sampled to a dimension of 256 by 256 pixels, chosen experimentally in order to reduce the computational costs without worsening the performances of the system. They will be referred as the *Original images*. An observation to be made is that the algorithms described in the following are applied just to the full lung area obtained by means of the segmentation algorithm described in chapter 1.

2.3 Enhancing the conspicuity and visibility of the nodules

As explained in the Introduction, the conspicuity of the nodules changes according to their size and their location in the lungs. An enhancement procedure with the aim of increasing the visibility of nodules with different conspicuity should be designed in order to eliminate all the various factors that lower their visibility, by means of highlighting their common characteristics.
The nodules are generally characterized by a circular shape and the higher grey values of their pixels, with respect to the ones in their surrounding. This is the reason why the multi-scale scheme realized starts by producing several smoothed versions of the image; each one of them is then subtracted from the *Original* one. The *Smoothed images* are obtained by convolving the original image with gaussian filters whose standard deviation, σ, takes values in the range $2 - 12$, according to the minimum and maximum possible pixel size of the nodule radius in the image down-sampled to the size of 256×256 pixels. For each scale, σ, a *Difference image* is then obtained by subtracting from the *Original image* its smoothed version. In the result of the subtraction the details visible at the scale σ are enhanced, as can be noticed in Figure 2.1. Since the distribution of the grey levels in a nodule sub-image can be approximated by a gaussian, the result of subtracting to a nodule sub-image its smoothed version is usually an image with a positive peak in the central part of the nodule, and negative values in the neighborhood. Moreover the histogram of the *Difference image* shows that most of the pixels take negative values, meanwhile on the set of positive values a peak can always be identified. To identify all the details belonging to the scale σ, the pixels with a value bigger than the one corresponding to the peak are selected and a binary image is created to store their position; these pixels are the ones corresponding to the highest frequencies, i.e. the details, that can be identified at the scale σ. Summing up all the binary images obtained at different scales a final *Sum image* is created, two examples of which are shown in Figure 2.2. Note how this new method, employing a multi-scale approach to enhance the conspicuity of the nodules, works well also in the case of extremely subtle nodules.

2.4 Extracting the nodule candidates

In the *Sum image* the nodules often appear as regions with circular shape of different sizes, characterized by the highest grey levels at the center and surrounded by a much darker ring. Based on this observation the image is processed to look for a measure which helps in selecting the pixels corre-

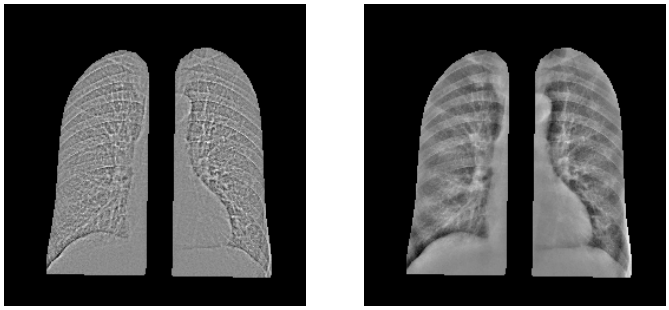

Figure 2.1: *Difference images* obtained with $\sigma = 2$, $\sigma = 12$

sponding to the centers of nodules. To handle all the possible sizes of the nodules the procedure described below is repeated for each possible radius value ($r = [2, .., 12]$) and all the results are then combined.
Having fixed the radius r, it is possible to calculate for each pixel $P = P(x, y)$ a coefficient $P_r(x, y)$ which measures the contrast between a circular region with center P (and radius r) and its surrounding; $P_r(x, y)$ is thus defined:

$$P_r(x, y) = MEAN(Circle_r(P(x, y))) - MEAN(Ring_r(P(x, y))) \quad (2.1)$$

where $MEAN(X)$ is the mean of the gray values of the pixels inside a generic region X; $Circle_r(P(x, y))$ is the region composed by the pixels contained in the circle of radius r and centered in P; $Ring_r(P(x, y))$ is the region composed by the pixels in the 2-pixel-thick ring around the $Circle_r(P(x, y))$. Note that the thickness of the ring is fixed to 2 for every radius. This is because what allows to identify a circular region is a darker ring surrounding it, no matter which is the thickness of the ring itself.
To select the pixels which are potential nodule centers, a threshold on the set of the coefficients $\{P_r(x, y)\}$ is automatically defined by means of the algorithm described in [42] and based on the maximization of the entropy of the histogram of the image: the threshold is set so that the two grey level subsets created have the maximum possible entropy. Thresholding the image with the value obtained a *Binary image* is created.
For each connected region in it, the circularity and the biggest diagonal, D, of the minimum ellipse containing the region itself are calculated. The circularity is defined as the fraction of the area of the region contained in the circle, with the same area and centered in the center of mass, and the area of the circle itself (the same circularity measure has been used also in [26]). The analysis of these values allows to discard a candidate region either if its circularity is lower than 0.5 or D is bigger than $2 * r$. The regions not discarded correspond to the candidate nodules with radius r. Repeating the procedure for each possible radius a set of 11 *Binary images* $B(r)$ is created, each containing a set of candidate nodules. All these images must be combined to extract a final set of candidates. First, all the regions

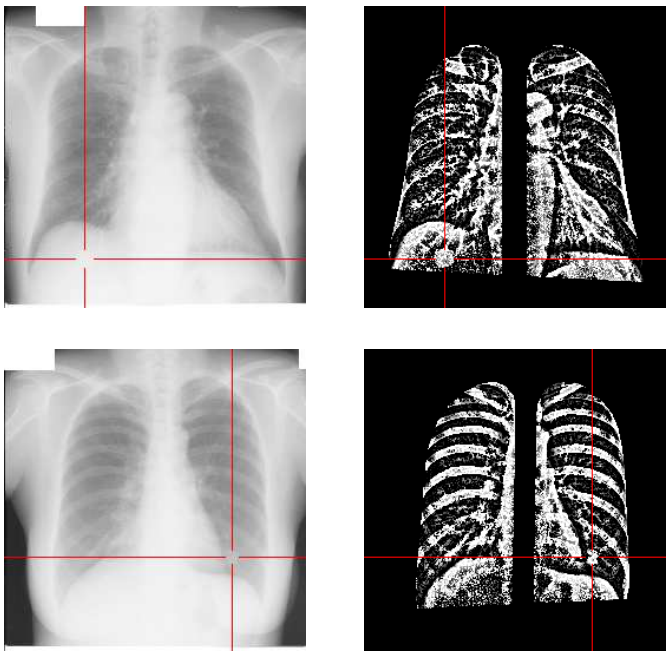

Figure 2.2: *Original image* and *Sum image* - subtle nodule behind the diaphragm and extremely subtle nodule

appearing in only one of the *Binary images* are taken as candidates. For the others the following procedure is employed: when two regions, X and Y, belonging to $B(r_1)$ and $B(r_2)$ (r_1 and r_2 being two consecutive radius values) intersect, their union U is at first considered. If the biggest diagonal D of the minimum ellipse containing it is less than $2 * r_2$, then U is taken as representative; otherwise the following mean values are calculated:

$$M_X = \frac{1}{\mid X \mid} \sum_{p \in X} P_{r_1} \quad and \quad M_Y = \frac{1}{\mid Y \mid} \sum_{p \in Y} P_{r_2} \tag{2.2}$$

and we take as representative region the one with the higher value.
A final grey level image containing all the candidate tumors, and called *Regions image*, is computed by assigning to each pixel in each candidate region the value

$$G(x, y) = \max_{r \in [2,12]} (P_r(x, y)) \tag{2.3}$$

and then scaling it in the range $[0, 255]$. Figure 2.3 and 2.4 show two examples of *Regions image* obtained for the case of a subtle and an extremely subtle nodule.

With this extraction scheme 31100 regions are obtained on all the 247 images of the $JSRT$ database, with an average of about 125 regions per

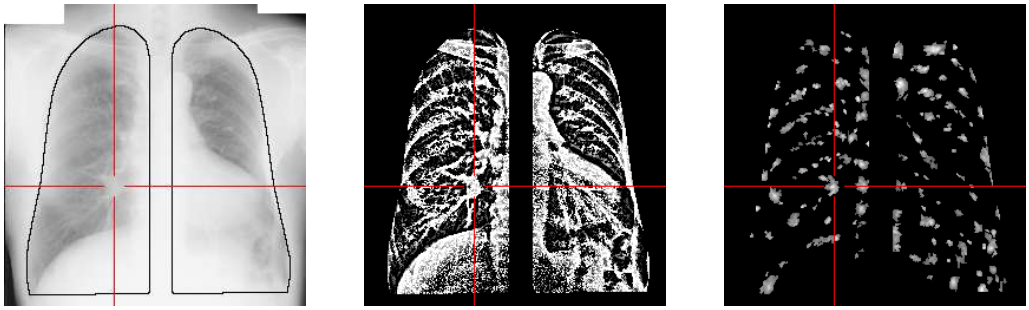

Figure 2.3: *Original Image, Sum and Regions Image* - subtle nodule

Figure 2.4: *Original Image, Sum and Regions Image* - extremely subtle nodule

image and only 5 True Positives lost out of 154. The images containing the lost tumors are shown in Figure 2.5: all of them have been classified in the *JSRT* Database as containing *extremely subtle* nodules; moreover showing them to radiologists, and indicating them the position of the nodule, as reported in the data of the database, they still doubt that the one in the second row on the left contains a nodule.
These results have been compared with those of the extraction schemes tested on the same database and reported in [1] and [8]. The first method is applied to the lung area defined by [85] but not extended as described in chapter 1, bringing to a loss of 13 True Positives, out of 154, even before the candidate extraction procedure. The result of the extraction scheme is a set of about 33000 candidates and a loss of other 8 True Positives; the final sensitivity ratio it reaches is then equal to 0.86. The implementation of the second method and its application to the same lung area used in [85], gave really poor results.
An observation regarding the number of candidates obtained at this first stage is that the lung area used by the scheme described in this chapter is about 1.5 times bigger than the one commonly considered; this number, that is already lower than the ones of the other schemes, would have been much less if we had used the area of the visible part only.

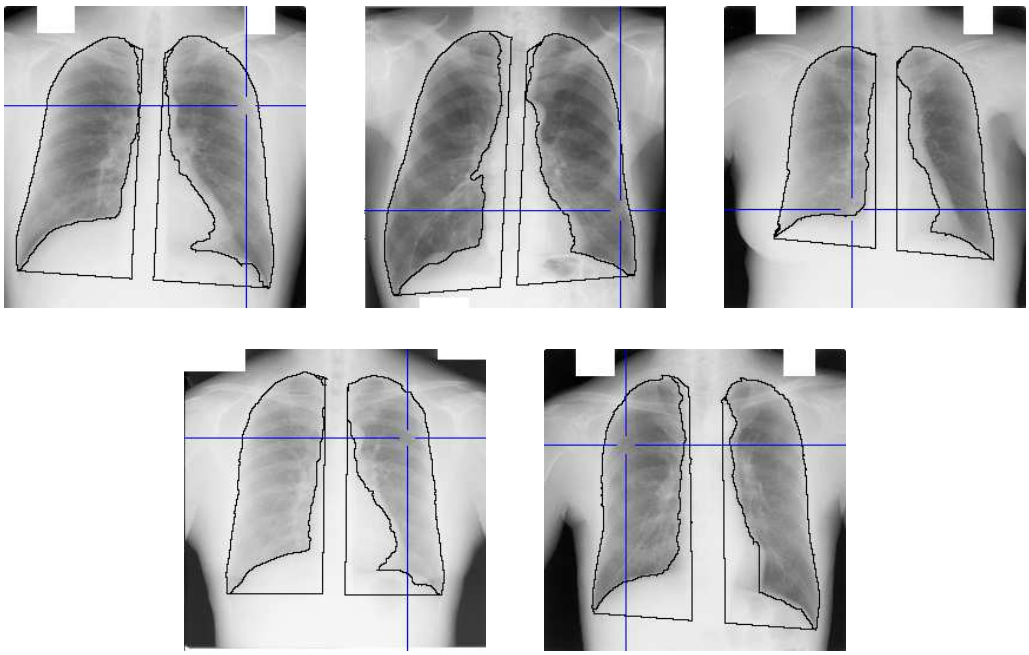

Figure 2.5: Radiographs containing nodules lost

2.5 Conclusions

In this chapter we have described a method to extract a first set of candidate nodules from the lung area defined by the segmentation algorithm presented in chapter 1; the sensitivity ratio obtained by the system is equal to 97%; it is higher than the other methods presented in the literature and tested on the same database. Despite this good performance the number of candidates selected, 31100 on 247 images (about 125 per image), is too high to be useful for clinical purposes and must be reduced. A good system should indeed get no more than $5 - 10$ candidates per image. In the next chapter the experiments aimed at a final True Positives selection are described in detail. When executed on a Pentium III with 128 Mb of RAM the algorithm takes a maximum of 20 seconds. This time can be reduced significantly, to at least its 10%, by programming the algorithm in a compiled language and optimizing the code.

Chapter 3

Experiments for Candidate Classification

In this chapter we describe the experiments aimed at the selection of the *True Positives* from the set of the extracted candidates, whose cardinality is too high to be useful for clinical purposes. The aim is to arrive at a final set containing no more than 10 candidates per image. To prune the candidate set, 16 most representative features have been calculated for each region and used as input to three different classifiers, whose performances have been compared. The classifiers are: a rule based system, Multi Layer Perceptrons (*MLPs*), and Support Vector Machines (*SVMs*).
The results obtained with all the classifiers, even if promising, are still not satisfactory and useful for clinical purposes. To improve the performance of the methods, different set of features have been computed and different types of SVMs have been trained using them as input. All the experiments and the results obtained are described with details.

3.1 Introduction

Several methods have been presented in the literature, which are aimed at *True Positive* selection from a first set of labelled candidate regions. Lots of them use a rule based system, where all the rules classify the regions using thresholds that are applied to set of features ([26], [28],[27], [56], [78], [77]). Other methods employ learning systems such as feed forward artificial neural networks or other learning machines ([51], [52], [95], [40], [63],[58], [47]). Xu et *al.* presented a classification method based on two consecutive steps: a rule based system is used to first reduce the big number of candidates; this classification is then followed by a learning algorithm. The same scheme has been used by Schilham and Ginneken in [1].
The basis of both a rule based system and a classification system is the selection of a proper set of features; this problem has been studied by McNitt-

Gray in [57]. Almost all features presented in the literature are classical features or their simple variations. Some of them describe the shape of the candidate region considered: examples are the circularity, the width, the radius of the circle including it; some features are used to study the texture of the sub-image containing the candidate and extracted from the original image: examples are simple features such as the mean, the standard deviation, the contrast of the grey values of the sub-image, but also more complicated ones such as the Laws' ([46]) and Haralick's ([31]) texture features. We noticed that there is also a considerable overlap between the features and the method employed in various studies.

In this chapter we describe our experiments whose purpose is the classification of the candidates in order to discard as many *False Positives* as possible without loosing too many *True Positives*. This task is made more difficult by the fact that the two sets to be separated are highly unbalanced.

The first problem we had to address is the choice of the features to be used as description of the candidates. At the state of the art, none of the methods computes features using the intermediate results created by the algorithms for the candidate extraction. This means that the way the regions are created and selected is not taken into account, even though it is obvious that the candidates obtained (*True Positives* and *False Positives*) are strongly dependent from them. This is the reason why some of the features used as input for our classification systems have been calculated from the images and values created and employed during the extraction scheme described in the previous chapter.

To lower the number of the candidates extracted a set of 16 most representative features has been calculated for each region. These features allowed us to create a rule based system composed of simple and intuitive rules, which is able to discard 22633 *False Positive* regions (thus 8467 candidates are left) without losing any *True Positives*. The Sensitivity ratio of the system is not reduced and the number of candidates is decreased to approximately 35 per image. The same feature set has been used as input to Neural Networks, trained by back propagation, and Support Vector Machines, and their performances have been compared with the ones obtained with the rule based system (see sections 3.4 and 3.5).

Furthermore the results obtained are compared with the ones of the system presented in the literature and tested on the same database ([1] and [8]), performing the two steps of classification previously described; the best performing between the two systems ([1]) employs a first step of classification that reaches a sensitivity ratio equal to 77% and a number of candidates equal to 5028; note that this last number is comparable to the one of our scheme since the lung area where the authors search for the candidates is equal to the 2/3 of the one we considered. The second step of classification of the system just mentioned is able to discard many other *False Positives*, but loosing to many *True Positives*.

Since the number of candidates left, by all the three methods experimented, is still too high to be useful for clinical purposes, we again tried to use Support Vector Machines but trained with more complex feature sets, i.e. composed of the 16 features already used plus more sophisticated ones.
The results, reported in the same paragraph (3.5), are worse than the ones obtained with the feature set of 16 elements. This fact suggested, as future works, to use feature selection techniques to select from big number of features a proper subset; as demonstrated before, a good combination of selected features could indeed be more discriminative than the whole set, thence allowing to obtain better results.

3.2 Features extracted to describe the Candidate regions

To prune the set of extracted candidates more than 40 features have been calculated. Their statistical distribution has been observed in order to derive a set of rules which could describe the main characteristics of the real nodules, therefore allowing to discard some *False Positives*. The statistical analysis allowed to select a set of 14 most representative features; their combination, by means of simple rules, has indeed proved to be effective for a first candidates selection. The rule based scheme created is able to detect and discard almost the 3/4 of the number of *False Positives*.
The features used are based on the shape and position of the regions, the grey level distribution in the original radiograph down sampled to the dimension of 512×512, the values of the grey levels in the *Regions image*, and the set of coefficients $P_r(x, y)$ associated to each pixel for each radius value (see section 2.4). The fact that some of the features are extracted with the values produced by the extraction scheme is a novelty with respect to the other methods in the literature, and it is due to the obvious and strong dependency between the regions obtained and the algorithm used to extract them.
This section describes how the selected features have been calculated; each one is associated with a simple name that will be used in the next section to make the description of the rules simpler.

Three features are calculated to describe the position of the region. The first position feature (called Lung Border Overlapping, or LBO in the following) has been introduced to eliminate false positives detected on the rib cage boundaries, which are characterized by the fact that they are attached to the lung borders and have an elongated shape. It is calculated considering the external contour of the region and computing the fraction of the pixels of the contour which lay outside the full lung area with respect to the total of the pixels in the contour. Figure 3.1 shows the elements used to calculate

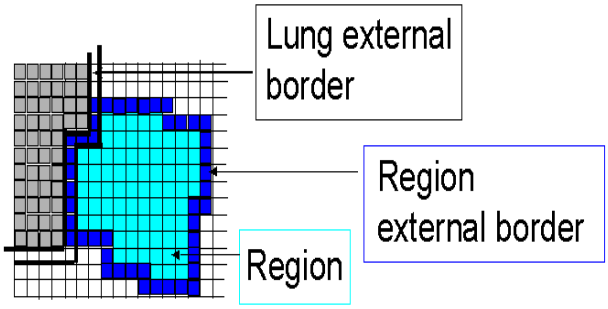

Figure 3.1: Lung Border Overlapping

this feature.

The other two features are calculated thanks to the information obtained with the segmentation aimed at defining a separation between the *visible lung areas* and the parts hidden by the spine, the heart and the diaphragm. One feature ($Position$) is indeed associated with a value corresponding to the location of the centroid of the region. It takes a value equal to 1 if the centroid is in the *visible lung area*, and equal to 0 otherwise; each candidate region is then considered as belonging to the lung area where its centroid is located. This last sentence is due to the fact that some regions may be partially located in the *visible lung part* and partially located in the *hidden parts*. These candidates are mainly false positives due to the high contrast at the borders of the two areas. To identify them another features is calculated, as an evaluation of the percentage of the region itself that lays outside the part of the lung where the centroid is located. This feature ($AreaOut$) is greater than zero just for the candidates crossing the borders between the *visible lung parts* and the *not visible lung parts*; it is calculated as the fraction of the number of pixels outside the lung part where the centroid of the region is located and the total number of pixels in the region itself. This quantity is a good evaluation of how much the region is on the border just mentioned and hence can be helpful to recognize the false positives there located.

Six features are based on the shape; they are:

- circularity ($Circ$), as defined in the previous chapter;
- effective radius ($EffRadius$) , that is the radius of the circle with an area equivalent to the one of the region;
- the fraction between the perimeter of the region and the perimeter of the circle with radius of 12 pixels ($NormalizedPerimeter$) - remember that the maximum radius value considered is 12 pixels-;
- the fraction between the area of the region and the area of the circle

with radius of 12 pixels ($NormalizedArea$);

- the fraction between the two dimensions (smallest/biggest) of the minimum bounding box ($MinBBFraction$);
- the fraction between the two dimensions (smallest/biggest) of the maximum bounding box ($MaxBBFraction$);

One feature is based on the mean grey level of the pixels of the region in the original radiograph down-sampled to the dimension of 512×512 ($MeanO$).

Two features are calculated on the grey level of the pixels in the *Regions image*: they are the mean ($MeanCR$) and the maximum value of the grey level of the pixels ($MaxCR$).

Two features are calculated as an estimate of the most characteristic radius value to be associated to a generic region X. Two different methods have been employed to get it and hence obtain two values that can be compared. One method first calculates for each pixel $P = P(x, y)$ a most eligible radius $rad(x, y)$. This is done by considering all the P_r calculated for that pixel and then calculating:

$$rad(x, y) = Argmax_{i\in[2,12]}(P_i(x, y)) \tag{3.1}$$

The first radius, R^1_X, associated with the region X is then equal to:

$$R^1_X = \max_{(x,y)\in X} rad(x, y) \tag{3.2}$$

The second method calculates the radius R^2_X by considering for each different value of r the sum

$$Sum_X(r) = \sum_{(x,y)\in X} P_r(x, y) \tag{3.3}$$

Again R^2_X is calculated according to:

$$R^2_X = Argmax_{i\in[2,12]}(Sum_X(r)) \tag{3.4}$$

The fact that R^1_X and R^2_X are very similar in case of true positive elements and significantly different for many false positives allows to discard some of them.

3.3 Rule Based System used to discard the False Positives

A simple rule based system has been developed based on the features previously described. The rules it employs have been formulated after the observation of the distribution of the single features or the relationships between

pairs of them. This system allows to discard 22633 false positive regions without loosing any true positive, hence keeping the sensitivity ratio equal to 0.97 and reaching a total number of candidates equal to 8467, approximately 35 candidates per image.

Some of the rules are based on thresholding of the input features. According to these rules a region is discarded if:

- $NormalizedArea > 1$
- $NormalizedPerimeter > 1.2$
- $EffRadius > 14$ OR $EffRadius < 2$
- $LBO > 0.35$
- $AreaOut > 0.3$

The rules describing the relationships between couple of features are listed below; some of them are also clarified with the use of some images. Figure 3.2 shows the relationship between the fraction of the two dimension of the maximum Bounding Box ($MaxBBFraction$) and the fraction of the two dimensions of the minimum Bounding Box ($MinBBFraction$). Each region is viewed as a point in the two dimensional space where $X = MaxBBFraction$ and $Y = MinBBFraction$. The points corresponding to the *True Positives* are always located in the band between the lines L_{Down} and L_{Up} which have been experimentally set; L_{Down} is the line passing by the points $P1 = (0.4, 0.35)$ and $P3 = (1., 0.6)$; L_{Up} is the line, parallel to L_{Down}, and passing by the point $P2 = (0.4, 0.6)$. The possible location is red colored in the corresponding image.

Figure 3.3 shows the relationship between the $LungBorderOveralapping$ (LBO) and the fraction of the two dimensions of the maximum Bounding Box ($MaxBBFraction$). This rule is based on the fact that a lot of *False Positive* candidates are located at the borders of the full lung area and have an elongated shape; the rule describes the requirement that regions nearby the borders cannot be too elongated. The line L_{Up}, passing by the points $P1 = (0, 0.3)$ and $P2 = (0.35, 0.95)$, is defined as a low boundary for the candidates; the regions corresponding to points located below this line are discarded.
The location of the not discarded regions is red colored in the corresponding image.

The boundary just defined, describing the relationship between LBO and $MaxBBFraction$, is used also to describe the relationship between the $AreaOut$ and $MaxBBFraction$. It is based on the fact that the *False Positives* detected due to the contrast in the border between the *visible lung parts*

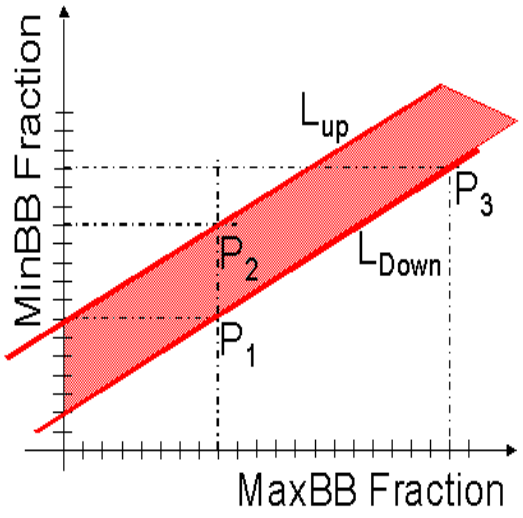

Figure 3.2: Relationship between $MaxBBFraction$ and $MinBBFraction$

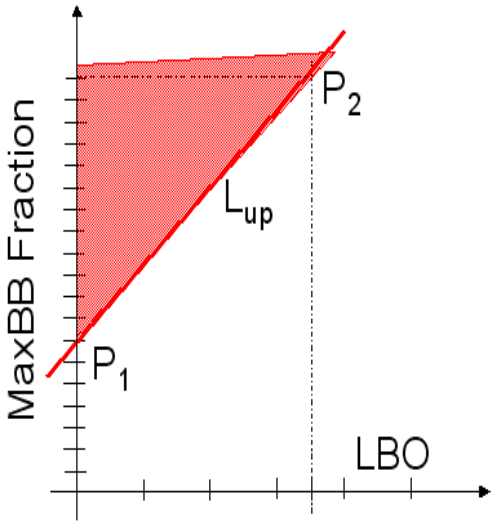

Figure 3.3: Relationship between LBO and $MaxBBFraction$

and the *hidden lung parts* are characterized by a bigger value of $AreaOut$, since their centroid is generally located nearby the border detected, and an elongated shape, i.e. a little value of $MaxBBFracction$.

The same reasoning can be used to explain the rules that relates LBO, and $AreaOut$, to the circularity $Circ$ of the region. The region is discarded if:

- ($LBO = 0$ AND $AreaOut = 0$) AND $Circ \leq 0.35$
- ($LBO > 0$ OR $AreaOut > 0$) AND $Circ \leq 0.55$

The following rules relate LBO to the mean grey level in the original Image ($MeanO$), and to the two features, $MeanCR$ and $MaxCR$, calculated with the values in the *Regions image*. Before applying the rules, $MeanCR$ and $MaxCR$ are normalized w.r.t. the maximum value in the *Regions Image*; $MeanO$ is normalized w.r.t. the maximum value in the *Original Image*. The rules defined are based on the fact that regions nearby the borders have brighter values both in the original image and in the *Regions image*. According to the rules a region is discarded if:

- $LBO = 0$ AND ($MeanCR \leq 0.45$ OR $MeanCR \geq 0.75$)
- $LBO = 0$ AND $MaxCR \leq 0.5$
- $LBO = 0$ AND $MeanO \leq 0.45$
- $LBO > 0$ AND ($MeanCR \leq 0.55$ OR $MeanCR \geq 0.75$)
- $LBO > 0$ AND $MaxCR \leq 0.75$
- $LBO > 0$ AND ($MeanO \leq 0.75$ OR $MeanO \geq 0.95$)

Regions located in the *hidden lung area* can be recognized by the feature $Position$ equal to 0; they are generally much brighter; this is the reason why their normalized feature $MeanO$ can be analyzed separately from the ones in the *visible lung area* to define a rule relating those two features. According to the rule defined a candidate is discarded if:

- $Position = 0$ AND ($MeanO \leq 0.75$ OR $MeanO \geq 0.95$)

The next rules express the relationship observed between the radius R_X^1 and the value $PFrac$ computed from the (not normalized) $Perimeter$ and the effective radius ($EffRadius$) as:

$$PFrac = Perimeter/(2\Pi\ EffRadius) \tag{3.5}$$

According to the rules defined a region is eliminated if:

- $R_X^1 < 7$ AND $PFrac \leq 1$ OR $PFrac \geq 1.5$

- $R_X^1 \leq 7$ AND $PFrac \leq 1.1$ OR $PFrac \geq 3$

The choice of the value of 7 is due to the fact that the radius takes values in the range $[2, .., 12]$; 7 is the mean value.

The last rules are aimed at describing the relationships between the three radius values calculated for each region. A candidate region is discarded if:

- $R_X^1 \geq EffR + 4$
- $EffRadius < 7$ AND $R_X^2 \geq 8$
- $EffRadius \geq 7$ AND $R_X^2 \geq 11$

In Figure 3.4 we show examples of the result of the pruning, realized by the rule based system, on patients with very subtle and extremely subtle nodules.
Our results can be compared with the ones of the method described in [1] in which the authors apply a classifier that selects 5028 candidates from the first set composed by 33000 regions, but loosing other 15 true positives in addition to the 21 already lost by the extraction method (as reported in section 2.4), for a total of 35 false negatives, thus decreasing the sensitivity ratio of the overall system to approximately 77%. Our pruning method is able to discard less false positives but does not loose any candidates. The robustness of the method is ensured by the fact that the rules chosen are very intuitive and simple. The drawback is the need of setting the thresholds, which may depend on the images in the database. This fact could be the cause of worse performances of the system when tested on images contained in other databases.

3.4 Neural Network Classifiers to discard the False Positives

In this section we describe the experiments aimed at the use of the feed-forward Neural Networks, trained by standard back-propagation, to trim the set of candidates extracted, composed by 31100 candidates. The advantage of such systems with respect to the rule based ones is the fact that, once the network is trained, it doesn't need any threshold to be set and, more important, it should be able to generalize with respect to unseen examples. Furthermore we expect that learning algorithms, working with the same set of features as input, could have a better performance since they could learn relationships between features that are more complicate than the simple ones of our rule based system.

We performed *Three* kind of experiments that differ for the vector that is

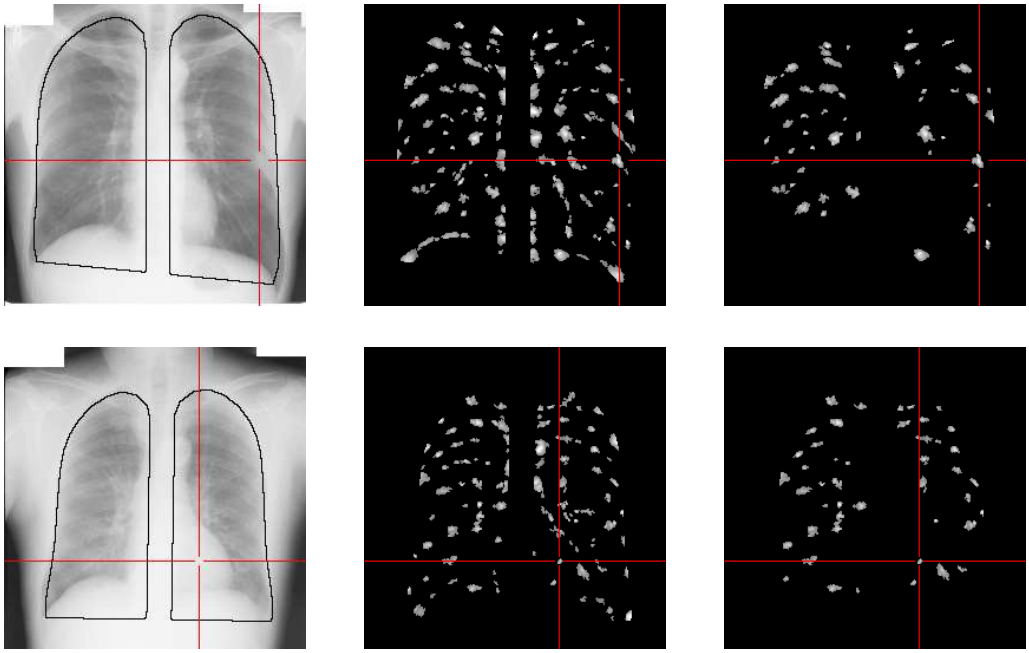

Figure 3.4: Original Image, Regions Image and Pruned set - very subtle nodule (top row) and extremely subtle nodule (bottom row)

chosen as input of the network, i.e. the representation of the data, and for the method used to create the training and validation sets. For each one of these three experiments we also tested various values of the network parameters and various architectures, to experimentally choose the best set-up of the learning system being trained.

The *first Two* experiments have been performed representing each candidate with one vector of 16 features $X = [x_1, ..., x_{14}, x_{15}, x_{16}]$, which is used as input of the network; specifically, $x_1, .., x_{14}$ represent the features previously described, and x_{15}, x_{16} are the spatial coordinates of the centroid of the region; they are expressed in a local coordinate system which has its origin in the center of mass of the lung field and it is scaled with respect to the width and length of the lung area. We apply to the input vector a preprocessing aimed to data normalization, followed by a scaling that brings all the values to the range $[0.0, 1.0]$.
These experiments were executed using both training and validation set.

In the *first* experiment we randomly split the available positive data in 75 examples for training (approximately the 50% of the total number), 15 examples for validation (approximately the 10% of 149) and the remaining 59 examples were used for testing. From the available large set of negative data we extracted without replacement a number of negative examples equal to

five times the number of the Positives, both for the training and the validation, obtaining respectively $75 \times 5 = 375$ negative examples for the training set and $15 \times 5 = 75$ negative examples for the validation set. The remaining elements were of course in the test set. Several experiments were run, using different network architectures, changing the parameters of the learning algorithm, and obviously the elements in the training, validation and test sets. During these experiments an input data was classified as belonging to a class when the corresponding output neuron had a value bigger than a threshold set to 0.5. In any case the results were really poor: lots of *True Positive* candidates were classified as *False Positive*, and this happened also after we lowered the threshold associated to the output neuron corresponding to the *Nodule* (i.e. *True Positive*) class. It is obvious that the network cannot discern the *True Positives* since their number is too little with respect to the *Not Nodules* in the training set; this is the motivation of the poor discriminative power of the MLP which classifies almost all the new element as belonging to the most represented class between the two learned.

According to the results obtained in the first experiment, the *second* one was performed, using the same input vector to represent the candidates, but in order to train the neural network to recognize and discard the regions whose set of features is totally different from the one of the nodules. To this end a new classification of the elements to be recognized was performed, obtaining three classes defined as:

- *"Real Nodules"*, containing the 149 True Positives;
- *"Possible nodules"* containing the 8318 False Positives not discarded by the rules;
- *"NOT nodules"* containing the 22633 elements discarded by the rule based system.

The training set of the network was composed as before by randomly selecting as Positive examples the 50% of the *"Real Nodules"* (75) plus the same percentage of elements from the class of the *"Possible Nodules"* (4160 elements), for a total of 4235 examples; the negative examples were extracted by random selection of the 50% of the elements in the *"NOT nodules"* class (11300 elements). Note that the training set created in this way, not only contains a bigger number of Positive examples, but it is also not so unbalanced as in the previous experiment. As before, the validation set was created by random selection of the 10% of the elements from the three classes, where the Positive examples were always the ones belonging to the first two. The remaining elements (approximately 40% of the total for each class) were used for testing. The neural network architecture, experimentally chosen, was composed by 1 hidden layer with 8 neurons and an output layer with 2 neurons; several experiments were performed changing the parameters of the learning algorithm (which are the momentum and the learning rate, the maximum number of epochs for the training, the minimum error allowed

on the training set) and especially the elements in the input sets. During these experiments an input data was classified as belonging to a class when the corresponding output neuron had a value bigger than 0.7. The results obtained are comparable with those of the rule based system: the network does not loose any true positives and in the worst case the number of false positives detected is never less than the 99% of the number recognized by the rule based system.

The *third*, and not successful, experiment was run by using as input to the network the sub image containing the candidate nodule. This sub image was extracted by the radiograph, down-sampled at a dimension of 1024×1024, by taking the maximum bounding box including the candidate region. Since all the sub images extracted had different sizes, they have been all down-sampled to a dimension of 8×8 pixels; this size has been chosen according to the minimum size of the sub images extracted. Several training experiments, using both training and validation set, were run.
These sets were formed as described in the *first* experiment, by choosing from the classes of the *"Nodules"*, containing 149 elements, approximately the 50% of the elements (i.e. 75 Positive examples), and from the class of the *"Not Nodules"* (30951 elements) a number of elements equal to $75 \times 5 = 375$. Again, 15 Positive and 75 Negative examples were randomly selected to create the validation set, meanwhile the remaining elements were used for testing.
During these experiments an input data was classified as belonging to a class when the corresponding output neuron had a value bigger than 0.5. Despite several architectures of the network have been experimented, and different settings of the learning parameters have been tried, no good results have been obtained: with the best settings we obtained a network that recognized, on the average, the 1/3 of the *True Positive* regions. Several reasons motivate this bad results. The highly unbalancing of the data has of course a negative effect. To this observation we may be add the fact that there are great differences between the sub images of the nodules, since too few of them are available; moreover their down-sampling causes a loss of information. To overcome these problems the best choice is indeed to use as input to the learning system a set of features describing each region.

3.5 Support Vector Machines to discard the False Positives

In this section the experiments aimed at the pruning of the set of extracted 31100 candidate regions, by means of Support Vector Machines, are explained and reported. This type of classifier has been widely used during the last years in a lot of different fields and for various classification tasks,

due to its good performances. A short introduction to Support Vector Machines is given in Appendix D.

Remember that the set of candidate regions considered is composed of 149 positive examples ("True Nodules"), and the remaining are negative examples ("False Positive"). Note that in the rest of this section the term False Positive is used to indicate negative examples classified as positive by the SVMs.
Three experiments are reported, which employ *Three* different data sets to train the classifiers.

3.5.1 Data sets

The *first* data set, called the *Original Data Set*, is the same one used as input to the Neural Networks (see section 3.4), and composed of 16 features.
The *second* data set, called *Geometric Data Set*, contains vectors of 110 *geometric features*.
The first 22 features are the 16 of the *Original Data Set*, plus other 6 features to better describe the shape and the position of the region; they are:

- the X-coordinate and Y-coordinate of the centroid of the region, expressed in the local coordinate system that has its origin in the center of mass of each lung (left or right, according to the one where the region is located) and it is scaled with respect to its width and length;
- the ratio between the area of the region and the area of the Maximum Bounding Box including it;
- the ratio between the area of the region and the area of the Minimum Bounding Box including it;
- the ratio between the area of the region and the area of the circle completely including it;
- the ratio between the Perimeter of the region and the perimeter of the circle with the same area of the region;

The other 88 features are calculated taking the mean, the standard deviation, the contrast and the maximum value of the grey levels of the pixels of the candidate region in 22 different images; these are the 22 images computed by the multi scale algorithm used to enhance the conspicuity of the nodules (see section 2.3 in chapter 2). Specifically the images used are the 11 *Smoothed version* of the original image, and the 11 *Difference images*. The contrast is calculated as the difference between the maximum and the minimum value of all the pixels in the region.

The *third* data set, called the *Gabor Data Set*, is composed by vectors of 88

features mainly obtained by means of the application of Gabor filters (for a description of Gabor filters, and how to create a Gabor filter bank, see Appendix E).
The first 22 features are the same described for the *Geometric Data set*.
The next 40 features are computed after filtering the sub image, containing the nodule candidate region and extracted from the original radiograph, with a bank of 40 Gabor filters. The filter bank is composed by Gabor filters at 5 different scales in 8 different directions. The scale parameter, k_ν, used in the function expressing the filter, takes certain takes the values 0.555, 0.392, 0.277, 0.196, 0.138; they are chosen so that the supports of the filters created for each scale have pixel sizes that cover the range of variation of all the possible pixel sizes of the nodules in the original image. (The supports of the filters created have pixel sizes which are equal to respectively 45, 64, 90, 128, 181; meanwhile the possible pixel size is in the range $[33, .., 193]$).
The 40 features are calculated by computing, for each filtered version of the sub image, the mean of the values of 9 pixels: they are the centroid of the region and its $8 - neighboring$ pixels. The next 10 features are calculated by splitting the 40 features just computed into 5 sets, each one containing all the features computed on the sub-images filtered with filters belonging to the same scale (in 8 different orientation). The 10 features calculated are the mean and standard deviation of the elements in each one of the 5 sets.
In the same way we have split the 40 features in 8 sets, each one containing the features calculated on the sub-images filtered with Gabor filters oriented in the same direction; thence, they are each composed of 5 values; computing the mean and the standard deviation of each one of the 8 sets considered, the last 16 features for the *Gabor Data Set* are calculated.

Considering that all the data sets are very unbalanced, for training and testing positive-enriched data sets were built. In particular Positive and Negative examples were considered separately. We randomly split the available positive data in 89 examples for training and 60 examples for testing according to a train/test ratio equal to 3/2. From the available large set of negative data we extracted without replacement a number of negative examples equal to five times the number of positive data, both for the training and the test set, obtaining respectively $89 \times 5 = 445$ negative examples for the training set and $60 \times 5 = 300$ negative examples for the test set. We randomly repeated the above process 10 times, obtaining 10 pairs of training and test sets. In all the experiments we normalized the components of the data vectors to 0 mean and unitary standard deviation.

3.5.2 Classification tasks

Three separated classification tasks were performed, using the *Original Data Set*, *Geometric Data Set* and the *Gabor Data Set*. In any case we applied

SVMs with linear, polynomial and gaussian kernels, varying the regularization parameter (in the literature usually referred to as the C parameter) between 0.001 and 1000, the degree in polynomial kernels between 2 and 6 and the "width" (σ parameter) in gaussian kernels between 0.01 and 10000. We computed the mean error and standard deviation on the training and test sets (considering the ten pairs of training and test sets).

In this context we need to obtain a high sensitivity in order to detect all the positive examples, without a significant loss in specificity, because from a medical point of view it is crucial to detect all the Positive examples, but at the same time we need to significantly reduce the number of False Positives. As explained before, if the number of False Positive is too high the Computer Aided Diagnosis becomes not useful and impractical, because the physician need to examine too many nodule images. For these reasons we computed the sensitivity, specificity and precision (see Appendix C for details).
With SVMs, lowering the decision threshold, we may increment the sensitivity at the possible expense of a decreased specificity. In this way we computed the specificity of the classifiers when the sensitivity is 1, that is we computed the fraction of False Positive when all the true positive nodules were correctly classified. Moreover, to better understand the behaviour of the classifiers, we performed a *ROC* analysis, to jointly evaluate in a synthetic way the sensitivity and specificity of the SVMs.

3.5.3 Results obtained with SVMs classification

Analysis of the overall test error

Tab. 3.1 shows the best results obtained with SVMs trained on the *Original Data Set*. The first columns reports the characteristics of the SVM: the type of the kernel and the corresponding regularization and kernel parameters (in polynomial kernel d stands for degree, in gaussian kernel σ represents the "width" of the kernel). In the second and third column are represented the test error averaged on 10 different test sets and the corresponding standard deviation, in the fourth and fifth column the train error and its standard deviation and the two last columns represent the sensitivity and the specificity.
Even if the average test error seems to be not too bad, we may see that the corresponding sensitivity (whose level is critical for this task) is very low (between 0.39 and 0.49 in the models with the lowest test error). If we add more geometric features, quite surprisingly, the results are worse: tab. 3.2 shows the best results obtained with SVMs trained on the *Geometric Data Set*. Indeed the estimated test error is significantly larger, and the sensitivity is lowered between 0.1 and 0.3. Similar results are obtained also with the *Gabor Data Set* (Tab. 3.3): a not negligible average test error and a very

Table 3.1: *Original Data Set.* Best SVM models with respect to the average test error.

SVM kernel	Test	Stdev	Train	stdev	Sens.	Spec.
Poly $d = 6$, $C = 2$	0.1181	0.0100	0.0940	0.0090	0.4550	0.9673
Linear $C = 0.1$	0.1194	0.0227	0.1174	0.0043	0.4444	0.9678
Poly $d = 5$, $C = 5$	0.1203	0.0059	0.0886	0.0098	0.4733	0.9610
Poly $d = 5$, $C = 2$	0.1203	0.0118	0.1015	0.0083	0.4350	0.9687
Poly $d = 2$, $C = 5$	0.1206	0.0067	0.1148	0.0130	0.4117	0.9730
Poly $d = 3$, $C = 5$	0.1206	0.0100	0.0970	0.0092	0.4583	0.9637
Gauss. $\sigma = 1000$, $C = 100$	0.1206	0.0107	0.1157	0.0113	0.4500	0.9653
Poly $d = 3$, $C = 20$	0.1208	0.0073	0.0886	0.0078	0.4883	0.9573
Poly $d = 3$, $C = 10$	0.1208	0.0080	0.0983	0.0108	0.4667	0.9617
Gauss. $\sigma = 1000$, $C = 1000$	0.1214	0.0085	0.0989	0.0103	0.4850	0.9573
Poly $d = 4$, $C = 1$	0.1214	0.0101	0.1163	0.0147	0.3850	0.9773
Poly $d = 5$, $C = 1$	0.1214	0.0109	0.1103	0.0088	0.4017	0.9740
Poly $d = 2$, $C = 20$	0.1217	0.0085	0.1017	0.0104	0.4667	0.9607
Poly $d = 3$, $C = 2$	0.1217	0.0086	0.1144	0.0140	0.3967	0.9747
Poly $d = 6$, $C = 1$	0.1217	0.0101	0.1051	0.0087	0.4133	0.9713

bad sensitivity is registered.
Note that it is not difficult to obtain no error with the training sets, both with the *Gabor Data Set* (Tab. 3.3), with the *Geometric Data Set* and with the *Original Data Set* (in these last two cases results with no errors are not shown in the related tables since not significant). This happens also when using gaussian or polynomial kernels trained on the entire data set (31100 examples): in this case we can easily achieve no error (that is sensitivity and specificity equal to 1), even if this fact does not guarantee generalization capabilities of the trained learning machines. SVMs are "strong" classifiers and can easily fit training data, but in this case generalization is not guaranteed.

Ranking of the SVMs with respect to the sensitivity.

Considering that sensitivity is so bad independently of the kernel used, we ranked the SVM models with respect to the sensitivity level (Tab. 3.4, 3.5 and 3.6). In the best case, with the *Original Data Set* we achieved a value of 0.53, with a slight decrement in the specificity (Tab.3.4). Unfortunately this value is largely insufficient to be useful for diagnostic purposes.
With the *Geometric Data Set* we achieved a value of 0.43, at the expense of a decrement in the specificity (Tab.3.5), and with the *Gabor Data Set* the results are slightly worse: sensitivity is about 0.37 in the best case (Tab. 3.6).

Ranking the real-valued outputs of the SVMs and ROC analysis.

In order to understand the reasons why the SVMs fail to separate positive from negative examples, we analyzed the real-valued discriminant function computed by the SVMs. Indeed we may use the real-valued discriminant

Table 3.2: *Geometric Data Set.* Best SVM models with respect to the average test error.

SVM kernel	Test	Stdev	Train	stdev	Sens.	Spec.
Gauss. $\sigma = 1000$. $C = 100$	0.1511	0.0112	0.0897	0.0117	0.2915	0.9603
Linear $C = 0.5$	0.1514	0.0082	0.1282	0.0121	0.2119	0.9759
Gauss. $\sigma = 10000$. $C = 1000$	0.1514	0.0097	0.1226	0.0143	0.2220	0.9739
Linear $C = 1$	0.1523	0.0073	0.1249	0.0123	0.2390	0.9695
Poly $d = 2$, $C = 2$	0.1531	0.0132	0.1015	0.0115	0.2492	0.9664
Linear $C = 2$	0.1531	0.0134	0.1193	0.0137	0.2492	0.9664
Poly $d = 2$, $C = 2$	0.1540	0.0120	0.1182	0.0129	0.1966	0.9759
Poly $d = 3$, $C = 1$	0.1545	0.0176	0.0646	0.0080	0.3119	0.9522
Poly $d = 3$, $C = 0.1$	0.1554	0.0082	0.1330	0.0092	0.1203	0.9895
Poly $d = 4$, $C = 0.1$	0.1556	0.0150	0.1006	0.0086	0.2051	0.9722
Poly $d = 2$, $C = 5$	0.1568	0.0125	0.0766	0.0112	0.3102	0.9498
Linear $C = 5$	0.1568	0.0145	0.1161	0.0116	0.2746	0.9569
Poly $d = 5$, $C = 0.1$	0.1573	0.0212	0.0684	0.0090	0.2593	0.9593
Linear $C = 0.1$	0.1579	0.0079	0.1420	0.0186	0.1322	0.9841
Gauss. $\sigma = 100$. $C = 1$	0.1582	0.0076	0.1287	0.0077	0.0966	0.9908

function computed by the SVM to rank the outputs of the SVM on the test set data. In this way lowering the threshold of the corresponding decision function we may increment the sensitivity at the expense of a lower specificity. To this purpose we ranked the real-valued outputs computed by the SVMs; then we lowered the decision threshold to correctly classify all the positive examples (that is to obtain a sensitivity equal to 1), and then we computed the corresponding specificity. Unfortunately the specificity obtained when the threshold is lowered to obtain a sensitivity equal to 1 is very low: about 0.11 on the average (Tab. 3.7) in the best case, and in most cases it also lower.

It is unlikely that these results are due to a weak discrimination capability of the SVMs. Indeed the threshold we need to correctly classify all the positive examples in most cases is very low (Tab. 3.7, columns $4, 5$ and 6): the ranking of many positive examples is very low, or, in other words, it seems that the SVMs "strongly believe" that many positive examples are negative. This may due to an "intrinsic ambiguity" of the data: examples classified as positive or negative, may not significantly differ with respect to the extracted features. The bad performances of the SVMs are summarized also in the ROC curves shown in Fig. 3.5 and 3.6. Fig. 3.5 shows the ROC curves obtained with a gaussian SVM (the second one in Tab. 3.7) applied to 10 different test sets drawn from the *Geometric Data Set*: the curves confirm the bad performances of the SVM. Similar results are obtained also with the *Gabor Data Set* (Fig. 3.6).

Table 3.3: *Gabor Data Set.* Best SVM models with respect to the average test error.

SVM kernel	Test	Stdev	Train	stdev	Sens.	Spec.
Gauss. $\sigma = 100$. $C = 2$	0.1553	0.0112	0.0605	0.0112	0.1659	0.9805
Gauss. $\sigma = 100$. $C = 5$	0.1568	0.0147	0.0137	0.0039	0.2614	0.9595
Poly $d = 4$, $C = 0.1$	0.1591	0.0159	0.0425	0.0073	0.2205	0.9650
Poly $d = 6$, $C = 0.1$	0.1591	0.0182	0.0000	0.0000	0.2455	0.9600
Gauss. $\sigma = 100$. $C = 10$	0.1610	0.0148	0.0028	0.0020	0.2955	0.9477
Poly $d = 5$, $C = 0.1$	0.1614	0.0154	0.0114	0.0032	0.2386	0.9586
Poly $d = 6$, $C = 1$	0.1614	0.0203	0.0000	0.0000	0.2455	0.9573
Poly $d = 6$, $C = 10$	0.1614	0.0203	0.0000	0.0000	0.2455	0.9573
Poly $d = 6$, $C = 100$	0.1614	0.0203	0.0000	0.0000	0.2455	0.9573
Poly $d = 6$, $C = 1000$	0.1614	0.0203	0.0000	0.0000	0.2455	0.9573
Poly $d = 6$, $C = 2$	0.1614	0.0203	0.0000	0.0000	0.2455	0.9573
Poly $d = 6$, $C = 20$	0.1614	0.0203	0.0000	0.0000	0.2455	0.9573
Poly $d = 6$, $C = 200$	0.1614	0.0203	0.0000	0.0000	0.2455	0.9573
Poly $d = 6$, $C = 5$	0.1614	0.0203	0.0000	0.0000	0.2455	0.9573
Poly $d = 5$, $C = 1$	0.1629	0.0227	0.0000	0.0000	0.2773	0.9491

Table 3.4: *Original Data Set.* Best SVM models with respect to the sensitivity.

SVM kernel	Sens.	Spec.	Test	Stdev	Train	stdev
Poly $d = 3$, $C = 200$	0.5333	0.9273	0.1383	0.0133	0.0579	0.0059
Poly $d = 2$, $C = 1000$	0.5317	0.9237	0.1417	0.0093	0.0607	0.0061
Poly $d = 4$, $C = 1000$	0.5317	0.8807	0.1775	0.0156	0.0139	0.0051
Gauss. $\sigma = 100$. $C = 1000$	0.5300	0.9003	0.1614	0.0195	0.0328	0.0080
Poly $d = 6$, $C = 200$	0.5300	0.8897	0.1703	0.0126	0.0150	0.0047
Poly $d = 5$, $C = 1000$	0.5300	0.8740	0.1833	0.0143	0.0047	0.0032
Poly $d = 5$, $C = 100$	0.5283	0.9083	0.1550	0.0191	0.0352	0.0071
Poly $d = 3$, $C = 100$	0.5267	0.9373	0.1311	0.0102	0.0642	0.0073
Poly $d = 2$, $C = 200$	0.5250	0.9407	0.1286	0.0086	0.0777	0.0069
Gauss. $\sigma = 100$. $C = 200$	0.5250	0.9357	0.1328	0.0114	0.0607	0.0071
Poly $d = 4$, $C = 200$	0.5250	0.9087	0.1553	0.0184	0.0399	0.0076
Gauss. $\sigma = 1000$. $C = 10000$	0.5233	0.9397	0.1297	0.0078	0.0766	0.0082
Poly $d = 4$, $C = 100$	0.5233	0.9220	0.1444	0.0182	0.0506	0.0068
Poly $d = 3$, $C = 1000$	0.5217	0.8967	0.1658	0.0165	0.0315	0.0051
Poly $d = 6$, $C = 1000$	0.5217	0.8687	0.1892	0.0106	0.0007	0.0010

Using very unbalanced training data.

In the previous examples we used positive-enriched data to train the SVMs. In this way we may obtain a better balance between positive and negative examples, but the cardinality of the training data is strongly reduced. In the experiments presented in this subsection we try to improve the cardinality of the training data at the expense of a larger unbalance between positive and negative examples.

We present here the results only with the *Original Data Set*, as in the previous experiments we obtained the best results with this data set. We considered separately positive and negative examples. Similarly to the previous experiments, we randomly split the available positive data in 89 examples

Table 3.5: *Geometric Data Set.* Best SVM models with respect to the sensitivity.

SVM kernel	Sens.	Spec.	Test	Stdev	Train	stdev
Poly $d = 2$, $C = 200$	0.4356	0.8749	0.1983	0.0175	0.0002	0.0006
Poly $d = 2$, $C = 1000$	0.4305	0.8749	0.1992	0.0182	0.0000	0.0000
Poly $d = 2$, $C = 100$	0.4237	0.8875	0.1898	0.0173	0.0054	0.0030
Gauss. $\sigma = 1000$. $C = 10000$	0.4068	0.8797	0.1992	0.0243	0.0000	0.0000
Poly $d = 3$, $C = 20$	0.4017	0.8895	0.1918	0.0224	0.0015	0.0015
Poly $d = 3$, $C = 200$	0.3932	0.8864	0.1958	0.0259	0.0000	0.0000
Poly $d = 2$, $C = 1000$	0.3932	0.8864	0.1958	0.0259	0.0000	0.0000
Poly $d = 3$, $C = 100$	0.3932	0.8864	0.1958	0.0259	0.0000	0.0000
Poly $d = 3$, $C = 10$	0.3847	0.8993	0.1864	0.0241	0.0098	0.0034
Poly $d = 3$, $C = 5$	0.3831	0.9142	0.1743	0.0221	0.0259	0.0057
Linear $C = 1000$	0.3831	0.8837	0.1997	0.0510	0.1100	0.0663
Poly $d = 2$, $C = 20$	0.3729	0.9207	0.1706	0.0226	0.0385	0.0081
Poly $d = 4$, $C = 20$	0.3729	0.8868	0.1989	0.0314	0.0000	0.0000
Poly $d = 4$, $C = 200$	0.3712	0.8864	0.1994	0.0327	0.0000	0.0000
Poly $d = 4$, $C = 1000$	0.3712	0.8864	0.1994	0.0327	0.0000	0.0000

Table 3.6: *Gabor Data Set.* Best SVM models with respect to the sensitivity.

SVM kernel	Sens.	Spec.	Test	Stdev	Train	stdev
Poly $d = 2$, $C = 10$	0.3727	0.8909	0.1955	0.0152	0.0011	0.0011
Poly $d = 2$, $C = 200$	0.3682	0.8859	0.2004	0.0149	0.0000	0.0000
Poly $d = 2$, $C = 20$	0.3682	0.8859	0.2004	0.0149	0.0000	0.0000
Poly $d = 2$, $C = 1000$	0.3682	0.8859	0.2004	0.0149	0.0000	0.0000
Poly $d = 2$, $C = 100$	0.3682	0.8859	0.2004	0.0149	0.0000	0.0000
Gauss. $\sigma = 1000$. $C = 10000$	0.3682	0.8782	0.2068	0.0130	0.0000	0.0000
Poly $d = 2$, $C = 5$	0.3636	0.8977	0.1913	0.0199	0.0075	0.0028
Gauss. $\sigma = 1000$. $C = 1000$	0.3636	0.8818	0.2045	0.0137	0.0013	0.0013
Poly $d = 3$, $C = 5$	0.3455	0.9109	0.1833	0.0192	0.0000	0.0000
Poly $d = 3$, $C = 200$	0.3455	0.9109	0.1833	0.0192	0.0000	0.0000
Poly $d = 3$, $C = 20$	0.3455	0.9109	0.1833	0.0192	0.0000	0.0000
Poly $d = 3$, $C = 1000$	0.3455	0.9109	0.1833	0.0192	0.0000	0.0000
Poly $d = 3$, $C = 100$	0.3455	0.9109	0.1833	0.0192	0.0000	0.0000
Poly $d = 3$, $C = 10$	0.3455	0.9109	0.1833	0.0192	0.0000	0.0000
Poly $d = 3$, $C = 2$	0.3386	0.9145	0.1814	0.0210	0.0005	0.0011

for training and 60 examples for testing according to a train/test ratio equal to 3/2. From the available large set of negative data we extracted without replacement a number of negative examples equal to 30 times the number of positive data, both for the training and the test set, obtaining respectively $89 \times 30 = 2670$ negative examples for the training set and $60 \times 30 = 1800$ negative examples for the test set. We randomly repeated the above process 10 times, obtaining 10 pairs of training and test sets. In all the experiments we normalized the components of the data vectors to 0 mean and unitary standard deviation. The results of Tab. 3.8 show that also this approach is unsuccessful: even if the SVMs can learn from more examples, the overall training set is probably too unbalanced. Indeed we achieved a very low sensitivity (0.32 in the best case). Of course the test error is lower, but this is

Table 3.7: SVM specificity when sensitivity is equal to 1. The first column refers to the obtained average specificity (computed on 10 different test sets), while the second and third column refer to the minimum and maximum specificity with respect to the 10 different test sets. The fourth column refers to the average value of the threshold we need to achieve a sensitivity equal to 1, while the fifth and sixth column report the minimum and maximum values of the threshold.

Specificity			Threshold			Data set and SVM kernel
average	min	max	average	min	max	
0.116818	0.0136364	0.281818	-1.00005	-1.00009	-1.00003	Gabor, Gauss. $\sigma = 10000, C = 0.001$
0.11661	0.0237288	0.267797	-1.00054	-1.00097	-1.00019	geometric, Gauss. $\sigma = 100, C = 0.001$
0.0915254	0.020339	0.186441	-1.59204	-2.22928	-1.21584	geometric, Gauss. $\sigma = 100', C = 10$
0.0831818	0.0590909	0.127273	-0.998548	-0.998893	-0.99825	Gabor, Gauss. $\sigma = 100, C = 0.01$
0.0818182	0.0136364	0.177273	-1.01111	-1.02057	-1.00287	Gabor, Gauss. $\sigma = 10000, C = 0.1$
0.0813636	0.0545455	0.122727	-0.986577	-0.990387	-0.983329	Gabor, Gauss. $\sigma = 10, C = 0.1$
0.0804545	0.0545455	0.127273	-0.777468	-0.841121	-0.723509	Gabor, Gauss. $\sigma = 10, C = 5$
0.0804545	0.0545455	0.127273	-0.777468	-0.841121	-0.723509	Gabor, Gauss. $\sigma = 10, C = 200$
0.0804545	0.0545455	0.127273	-0.777468	-0.841121	-0.723509	Gabor, Gauss. $\sigma = 10, C = 20$
0.0804545	0.0545455	0.127273	-0.777468	-0.841121	-0.723509	Gabor, Gauss. $\sigma = 10, C = 10000$
0.0804545	0.0545455	0.127273	-0.777468	-0.841121	-0.723509	Gabor, Gauss. $\sigma = 10, C = 1000$
0.0804545	0.0545455	0.127273	-0.777468	-0.841121	-0.723509	Gabor, Gauss. $\sigma = 10, C = 100$
0.0804545	0.0545455	0.127273	-0.777468	-0.841121	-0.723509	Gabor, Gauss. $\sigma = 10, C = 10$
0.0804545	0.0545455	0.127273	-0.777457	-0.841121	-0.723509	Gabor, Gauss. $\sigma = 10, C = 2$
0.0804545	0.0545455	0.122727	-0.866305	-0.903155	-0.833965	Gabor, Gauss. $\sigma = 10, C = 1$
0.08	0.00677966	0.254237	-1.00943	-1.02519	-1.0023	geometric, Poly $d = 3, C = 0.1$
0.08	0.00677966	0.183051	-1.0781	-1.16484	-1.02147	geometric, Gauss. $\sigma = 1000, C = 1$
0.079322	0.00677966	0.19661	-1.0079	-1.01688	-1.00195	geometric, Gauss. $\sigma = 1000, C = 0.1$
0.0762712	0.00677966	0.223729	-1.0957	-1.25254	-1.02274	geometric, Poly $d = 3, C = 0.1$
0.075	0.0318182	0.127273	-3.0195	-3.61612	-2.14827	Gabor, Gauss. $\sigma = 10000, C = 1000$
0.075	0.0136364	0.168182	-2.59202	-3.42032	-1.34429	Gabor, Gauss. $\sigma = 10000, C = 200$
0.075	0.0136364	0.163636	-1.05714	-1.1066	-1.00825	Gabor, Linear $C = 0.01$
0.0749153	0.0135593	0.162712	-1.34307	-1.58795	-1.19221	geometric, Gauss. $\sigma = 100, C = 1$
0.0749153	0.0101695	0.166102	-1.0374	-1.06857	-1.01933	geometric, Gauss. $\sigma = 100, C = 0.1$
0.0745455	0.0136364	0.172727	-1.05557	-1.10359	-1.00904	Gabor, Gauss. $\sigma = 10000, C = 5$
0.0740909	0.0136364	0.177273	-1.11092	-1.20662	-1.0175	Gabor, Gauss. $\sigma = 10000, C = 10$
0.0740909	0.0136364	0.172727	-2.06812	-2.81113	-1.17646	Gabor, Gauss. $\sigma = 10000, C = 100$
0.0731818	0.0136364	0.172727	-1.22179	-1.4142	-1.03424	Gabor, Gauss. $\sigma = 10000, C = 20$
0.0731818	0.0136364	0.163636	-1.57107	-2.06146	-1.08487	Gabor, Linear $C = 0.1$

due to the unbalance of the data.

A possible new approach

Since we deal with very unbalanced data, we may introduce a cost-sensitive approach to improve the sensitivity of the SVMs. Usually in classification problems the 0/1 loss function is applied: this loss function weighs equally errors on both positive and negative examples. In medical problems the cost of misclassifying positive (diseased) patients is usually higher than misclassifications of negative (healthy) patients. This is of course our case: we need to correctly classify all positive examples, possibly misclassifying only a low fraction of the negative ones. In the framework of SVM optimization problem we may introduce cost factors C_+ and C_- to be able to adjust the cost

Table 3.8: *Original Data Set* with very unbalanced data. Best SVM models with respect to the sensitivity.

SVM kernel	Sens.	Spec.	Test	Stdev	Train	stdev
Poly $d = 5$, $C = 1000$	0.3254	0.9771	0.0436	0.0044	0.0061	0.0008
Poly $d = 6$, $C = 200$	0.3119	0.9839	0.0374	0.0053	0.0109	0.0004
Poly $d = 6$, $C = 1000$	0.3119	0.9732	0.0478	0.0061	0.0037	0.0007
Poly $d = 4$, $C = 1000$	0.3017	0.9819	0.0397	0.0047	0.0109	0.0012
Poly $d = 5$, $C = 200$	0.2881	0.9878	0.0344	0.0034	0.0146	0.0010
Poly $d = 6$, $C = 100$	0.2746	0.9884	0.0342	0.0029	0.0145	0.0006
Gauss. $\sigma = 10$. $C = 10000$	0.2746	0.9769	0.0454	0.0031	0.0000	0.0000
Gauss. $\sigma = 10$. $C = 1000$	0.2746	0.9769	0.0454	0.0031	0.0000	0.0000
Gauss. $\sigma = 100$. $C = 1000$	0.2712	0.9917	0.0312	0.0023	0.0183	0.0010
Poly $d = 3$, $C = 1000$	0.2712	0.9894	0.0334	0.0037	0.0179	0.0012
Gauss. $\sigma = 10$. $C = 20$	0.2712	0.9888	0.0340	0.0021	0.0062	0.0010
Poly $d = 5$, $C = 100$	0.2678	0.9911	0.0318	0.0038	0.0186	0.0012
Poly $d = 4$, $C = 200$	0.2644	0.9921	0.0310	0.0030	0.0190	0.0014
Gauss. $\sigma = 10$. $C = 10$	0.2339	0.9930	0.0311	0.0020	0.0126	0.0008
Poly $d = 4$, $C = 100$	0.2169	0.9937	0.0310	0.0022	0.0223	0.0015

of false positives versus false negatives. Hence the minimization problem associated to SVMs is translated in the following one, where asymmetrical loss functions are used (see [44]):

$$\begin{array}{ll} \text{Minimize} & \mathbf{w} \cdot \mathbf{w} + \mathbf{C}_+ \sum_{\mathbf{i}:\mathbf{y_i}=+\mathbf{1}} \xi_\mathbf{i} + \mathbf{C}_- \sum_{\mathbf{j}:\mathbf{y_j}=-\mathbf{1}} \xi_\mathbf{j} \\ \text{subject to} & y_k(\mathbf{w} \cdot \mathbf{x_k} + \mathbf{b}) \geq \mathbf{1} - \xi_\mathbf{k} \\ & \xi_k \geq 0 \\ & 1 \leq k \leq n \end{array}$$

In the preliminary experiments presented in this section we applied this cost sensitive approach to the very unbalanced *Original Data Set* (in the training set we used a positive versus negative ratio equal to 1/30). In particular we considered cost-factors C_+ and C_- obtained by considering $C_- = C$ and $C_+ = C \times C_f$; we run experiments where C_f was set equal to 2, 5, 10, 20, 50, 100, so that training errors on positive examples outweigh errors on negative examples.
The results are summarized in Tab. 3.9. We achieved a significantly higher sensitivity with respect to the previous approaches. With relatively low values of C and quite large values of the cost factor C_f by which errors on positive examples are outweighed with respect to negative ones, we obtained sensitivity equal or larger than 0.90 and specificity equal about to 0.70. We can also obtain a sensitivity equal to 1, but at the expense of a very low specificity (data not shown).

These results are quite encouraging, even if probably not sufficient for clinical pre-screening of chest radiographs. Anyway ROC analysis of these results might offer insights to improve jointly sensitivity and specificity of cost-

sensitive SVMs. Moreover the better results obtained with the Original set of features, suggest that applying feature selection methods we could choose subsets of more informative features, both from geometric and Gabor ones.

Table 3.9: *Original Data Set* with very unbalanced data and using asymmetrical cost functions. Best SVM models with respect to the sensitivity, and with specificity larger than 0.50

SVM kernel	Sens.	Spec.	Test	Stdev	Train	stdev
Poly $d = 7$, $C = 0.001$, $C_f = 100$	0.9458	0.5566	0.4311	0.0334	0.4215	0.0352
Poly $d = 3$, $C = 0.01$, $C_f = 100$	0.9458	0.5249	0.4618	0.0369	0.4481	0.0370
Gauss $\sigma = 1000$, $C = 5$, $C_f = 100$	0.9356	0.5170	0.4697	0.0365	0.4617	0.0411
Poly $d = 4$, $C = 0.01$,$C_f = 100$	0.9288	0.5929	0.3964	0.0364	0.3864	0.0347
Gauss $\sigma = 1000$, $C = 10$, $C_f = 100$	0.9254	0.5364	0.4512	0.0351	0.4423	0.0384
Linear$d = 6$, $C = 0.01$, $C_f = 100$	0.9254	0.5062	0.4805	0.0376	0.4750	0.0364
Gauss $\sigma = 1000$, $C = 20$, $C_f = 100$	0.9220	0.5514	0.4368	0.0343	0.4234	0.0372
Gauss $\sigma = 1000$, $C = 100$, $C_f = 100$	0.9220	0.5191	0.4681	0.0301	0.4587	0.0316
Poly $d = 2$, $C = 0.1$, $C_f = 100$	0.9186	0.5920	0.3976	0.0401	0.3860	0.0375
Gauss $\sigma = 10000$, $C = 1000$, $C_f = 100$	0.9186	0.5538	0.4346	0.0324	0.4237	0.0372
Gauss $\sigma = 100$, $C = 1$, $C_f = 100$	0.9153	0.5919	0.3978	0.0359	0.3848	0.0352
Poly $d = 3$, $C = 0.001$, $C_f = 50$	0.9119	0.6086	0.3818	0.0335	0.3704	0.0394
Poly $d = 4$, $C = 0.001$, $C_f = 50$	0.9051	0.6490	0.3429	0.0266	0.3339	0.0337
Poly $d = 2$, $C = 0.1$, $C_f = 50$	0.8814	0.7007	0.2936	0.0174	0.2862	0.0158
Gauss $\sigma = 100$, $C = 2$, $C_f = 50$	0.8780	0.7202	0.2748	0.0201	0.2723	0.0162

3.6 Conclusions

In this chapter we described the experiments aimed at the selection of the *True Positives* from the set of the extracted candidates, whose cardinality is too high. Three different classification systems have been experimented, whose input is always a feature set composed by 16 most representative features. The systems used are a Rule Based system, the Multi Layer Perceptrons and the Support Vector Machines; the comparison of the results obtained show that the best performances are obtained by the Rule Based system. Nevertheless, the fact that the thresholds used by the rules are experimentally set doesn't ensure the robustness of this method with respect to a change of the database; this lead us to further experiments, whose purpose is the improvement of the results of the $SVMs$. Thus, more features have been experimented as input to $SVMs$ trained with different kernels and different value of the parameter C.
The *ROC* analysis has shown that promising results can be obtained when the training is performed by setting different, and very unbalanced, values for the cost factors C_+ and C_- and using as input of the training a very unbalanced data set selected from the one composed by the vectors of 16 features (see section 3.5.3). The fact that classifiers trained with this smaller

set of features have better performance than the ones trained with bigger ones, suggest as future development the use of feature selection techniques that could select, between all the features considered, a subset that could be more discriminative. Moreover the use of new features will be investigated, together with new classifiers or ensembles of them.

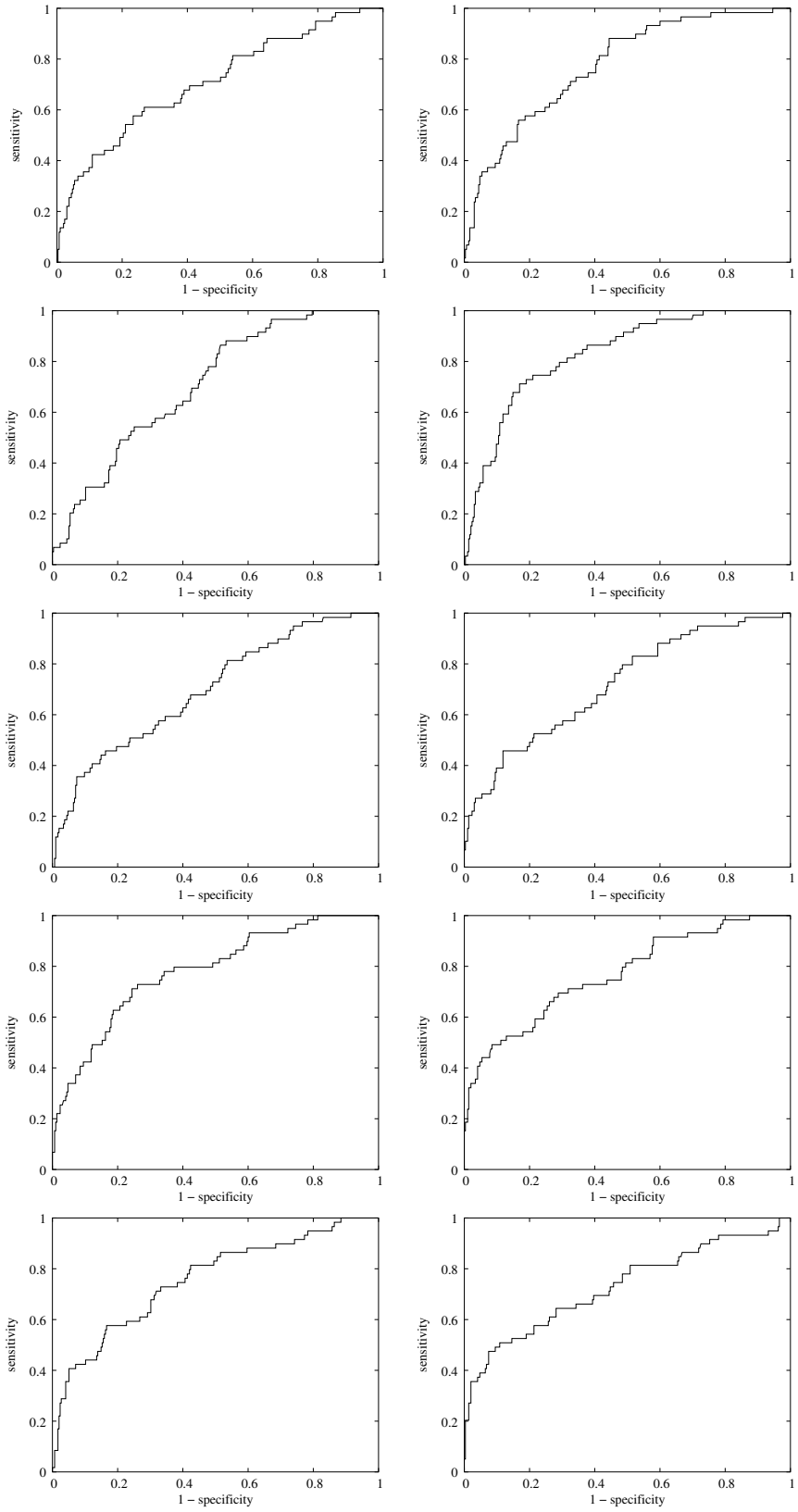

Figure 3.5: ROC curves obtained with the *Geometric Data Set*, using a gaussian kernel with $\sigma = 100$ and $C = 0.001$. The curves refer to 10 different test sets drawn from the *Geometric Data Set*.

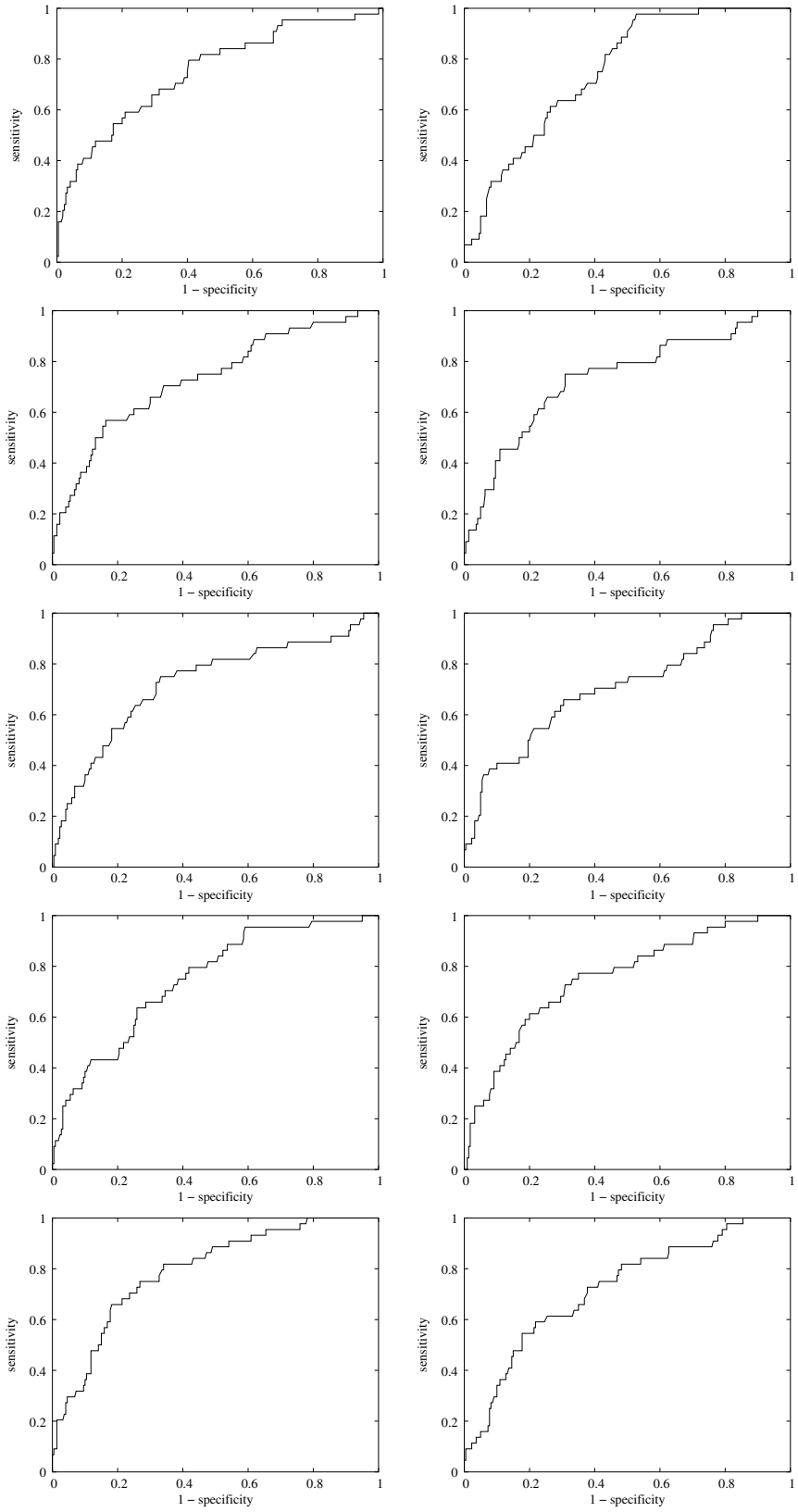

Figure 3.6: ROC curves obtained with the *Gabor Data Set*, using a gaussian kernel with $\sigma = 10000$ and $C = 0.001$. (the first SVM in Tab. 3.7). The curves refer to 10 different test sets drawn from the *Gabor Data Set*.

Summary, Conclusions and Future Works

In this thesis work an automatic system for tumors detection in Postero-Anterior chest radiographs is described; the final aim is the realization of a CAD system aimed at the early detection of this type of pathology.

The Introduction reports in detail all the reasons why the availability of such systems would be of utmost importance in the field of medical diagnosis on chest radiographs.

In chapter 1 a system aimed at lung area segmentation is described. It is a rule based method which employ a multi-scale technique. The result of this method is different from the ones created by the algorithms presented in the literature; it indeed includes in the segmented lung area also the parts of the lung that are hidden behind the spine, the diaphragm and the heart. This is because tumors may be present also there, and the segmentation is aimed at defining the area where the algorithms for their extraction will be applied. Nevertheless it also produces a good separation between the hidden part and the visible one, thus providing more information that can be used to better describe the nodule candidate regions extracted. A comparison with the methods described in the literature proved the better efficacy of the algorithm here presented; it indeed outperforms also the system ([85]), usually described as the best at the state of the art. The novelty of the system presented is the use of a multi-scale method, which allows to retrieve the correct edges of the thorax by exploiting the information belonging to different scales. This is because errors related to the edges found at a certain scale are easily detected and corrected by means of the information belonging to the other scales. The robustness of the method is ensured by the fact that it has been tested on 409 images belonging to 2 different databases created by different imaging systems, which employ different pixel resolutions and are subject to different types of quantum noise.

two consecutive multi-scale techniques, described in chapter 2, have been applied to the images (and restricted to the lung area defined with the segmentation algorithm) with the aim of producing a first enhancement of the

conspicuity of the nodules, followed by the extraction of a first set of candidate regions. Here the multi-scale method ensures that nodules of different size and conspicuity are detected. This system have also shown to have very good performance, especially if compared to the methods presented so far in the literature and tested on the same database ([1] and [8]).

The last step of candidate selection is the one that needs to be improved since the results obtained, even though promising, are still not sufficient for clinical purposes. This problem is common to all the schemes presented. So far no system has indeed been realized, whose performances allow to make it clinically available and used as a routine basis. All the experiments aimed at candidates selection are described in chapter 3. Both a rule based system and different learning machines ($MLPs$ and $SVMs$) have been experimented. The input of these systems is a data set composed by vectors of 16 most representative features to describe each candidate. The novelty of the features used relies in the fact that some of them are calculated from the intermediate images and measures, computed and used to perform the extraction of the candidates itself. Since the candidates extracted are obviously strongly dependent and related to these images and measures, it is straightforward to expect that features computed with them, could be helpful also to recognize the *True Positives* from the *False Positives*. The other features used are classical features describing the shape an the grey level characteristics of the candidates.
The rule based system, composed by simple and intuitive rules, obtains the best performance. Nevertheless, the thresholds used by the rules have been experimentally set on the overall set of features. This fact doesn't ensure the robustness of the method with respect to the change of the database. For this reason future works will be aimed at experimenting systems, such as the *Decision Trees*, that can be automatically created (or "grown") by means of a training set of data, in order to perform the same task of the rules.
The experiments employing $MLPs$ and $SVMs$ trained with the data set of 16 features, had performance that were not as good as the ones of the rule based system, but demonstrated the better efficacy of the $SVMs$ with respect to the $MLPs$. Furthermore, other experiments were aimed at training the Support vector machines with other two data sets composed by features vectors containing more elements; both the new data sets contain vectors composed of the same most representative 16 features plus others and more sophisticated. The results obtained are not better than the ones produced by the $SVMs$ trained with the original data set of the 16 features statistically selected; this means that more complicated set of features do not necessarily improve the performances of the classifier, hence suggesting the need of applying feature selection methods that could choose subsets of more informative features, from the big set composed of all the features considered.

Furthermore, we run other experiments using $SVMs$ trained with a very unbalanced data set, extracted from the original one; moreover we decided to set different, and very unbalanced, values of the cost factors C_+ and C_-, so that the training on positive examples outweighs errors on negative examples (see 3.5.3). With these settings better results have been obtained. This is a promising and encouraging first step. Anyway ROC analysis of these results might offer insights to improve jointly sensitivity and specificity of cost-sensitive $SVMs$. The use of *feature selection techniques* to form more informative data set to be used as input of $SVMs$ trained with the settings just mentioned could bring to an improvement of the performances. This approach could also be combined with *ensemble methods*, such as Random Subspace ensembles, in order to improve the accuracy and the reliability of the predictions.

The experiments presented, together with the systems and results that have been described in the literature, motivate the fact that nodules detection in chest radiographs remains an open and very difficult task in the field of Computer Aided Diagnosis, that is still far from being solved.
Nevertheless, given the widespread use and the clinical importance of this kind of radiographic analysis, it is likely that Computer Aided Diagnosis aimed at tumors detection in chest radiographs would be one of the first areas for commercial exploitation.

Appendix A

Oriented Derivatives of Gaussian Filters

Edges are usually defined as those points in an image where the gradient magnitude is maximum in the gradient direction [13]. Ridges are commonly defined as extrema in the direction of the largest curvature. Instead of these definitions, we consider structures defined by extrema in a *fixed* direction, for derivatives of a certain order in that same direction.
Derivatives are computed by convolving the image $I(x, y)$ with the derivative of a Gaussian $G(x, y; \sigma)$ at a particular scale σ. The normalized Gaussian in $2D$ is given by

$$G\left(x, y; \sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2+y^2/2\sigma^2)} \tag{A.1}$$

See Figure A.1 where a $2D$ gaussian function is drawn.
If we denote the nth order derivative in the direction defined by an angle

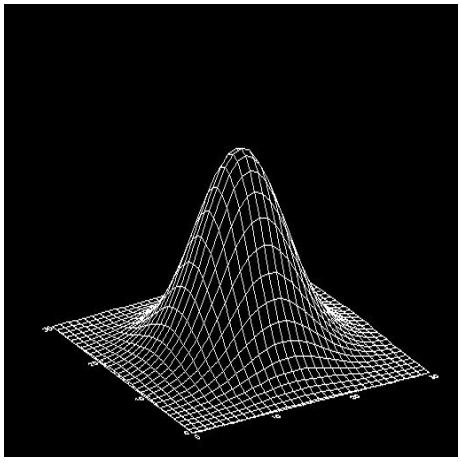

Figure A.1: *Bidimensional gaussian function*

α as I_n^α we need to compute:

$$I_n^\alpha(x, y; \sigma) = G_n^\alpha(x, y; \sigma) * I(x, y) \tag{A.2}$$

where $*$ denotes convolution, and G_n^α is the nth order derivative of the Gaussian kernel in the direction α. Instead of calculating the convolution directly, it is often convenient to express the response for an arbitrary direction α in terms of combinations of a finite set of basis filters (such constructions are commonly referred to as *steerable* filters). In particular in $2D$, the derivatives of any order n and for any direction α, can be computed from the set of all derivatives up to order n derivatives in two fixed orthogonal directions, e.g. x and y by substituting $x' = xcos\alpha - ysin\alpha$ and $y' = -xsin\alpha + ycos\alpha$. Other basis filters are also possible. See [24] for a general discussion on steerable filters. In particular, for order $n = 1$ and $n = 2$ one obtains:

$$I_1^\alpha(x, y; \sigma) = cos\alpha I_x + sin\alpha I_y \tag{A.3}$$

$$I_2^\alpha(x, y; \sigma) = cos^2\alpha I_{xx} + cos\alpha sin\alpha I_{xy} + sin^2\alpha I_{yy} \tag{A.4}$$

where, on the right-hand side, subscripts denote differentiation to x and y. More precisely I_x on the right hand side denotes the convolution of the image with the horizontal derivative of the gaussian filter; In the same way I_y denotes the convolution of the image with the vertical derivative of the gaussian filter.
Their mathematical formulations are:

$$I_x(x, y) = I(x, y) * G_x(x, y) = I(x, y) * \frac{d}{dx} G(x, y) \tag{A.5}$$

$$I_y(x, y) = I(x, y) * G_y(x, y) = I(x, y) * \frac{d}{dy} G(x, y) \tag{A.6}$$

In figure A.2 e A.3 $G_x(x, y)$ and $G_y(x, y)$ are shown.

The order (n) of derivative determines the nature of the detected structures. Using $n = 1$ yields edges from dark to bright regions, or vice versa; $n = 2$ finds bright and dark line structures, all in the direction perpendicular to α. An Example of the result of applying to an image the oriented derivative of gaussian filters in 8 directions is shown in Figure A.4. In Figure A.5 the binary images created by keeping the 1% of the pixels with the highest values are shown.

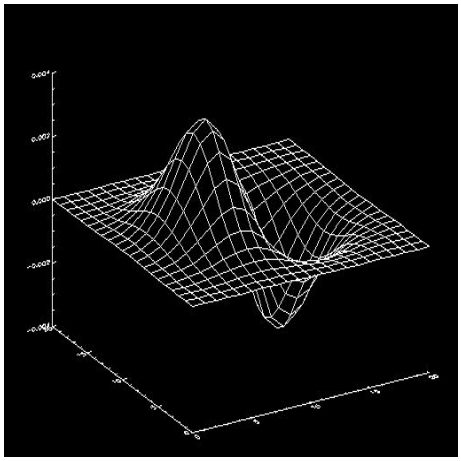

Figure A.2: *Horizontal derivative of the gaussian filter*

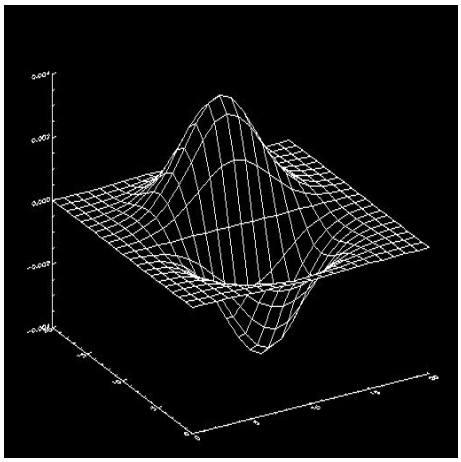

Figure A.3: *Vertical derivative of the gaussian filter*

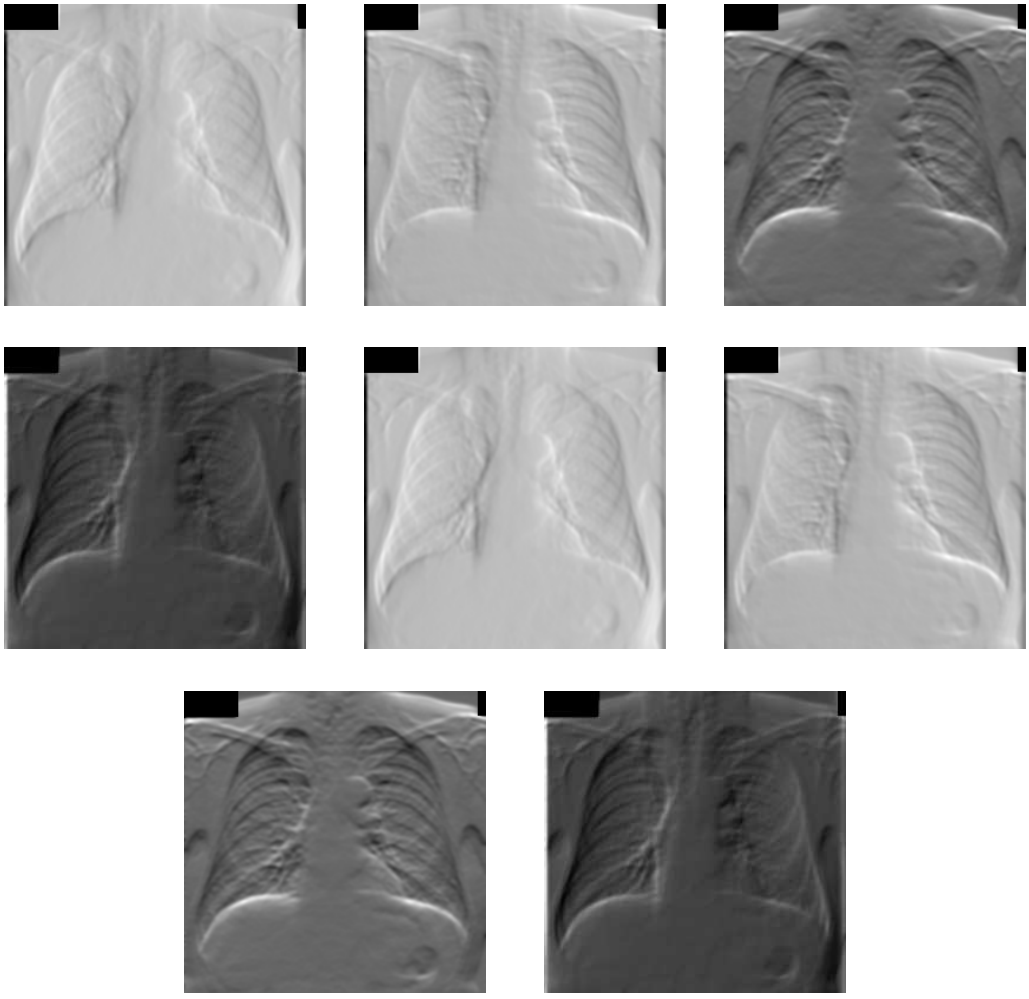

Figure A.4: *The result of the convolution of a chest image with a Gaussian derivative Filter oriented in* 8 *different directions.*

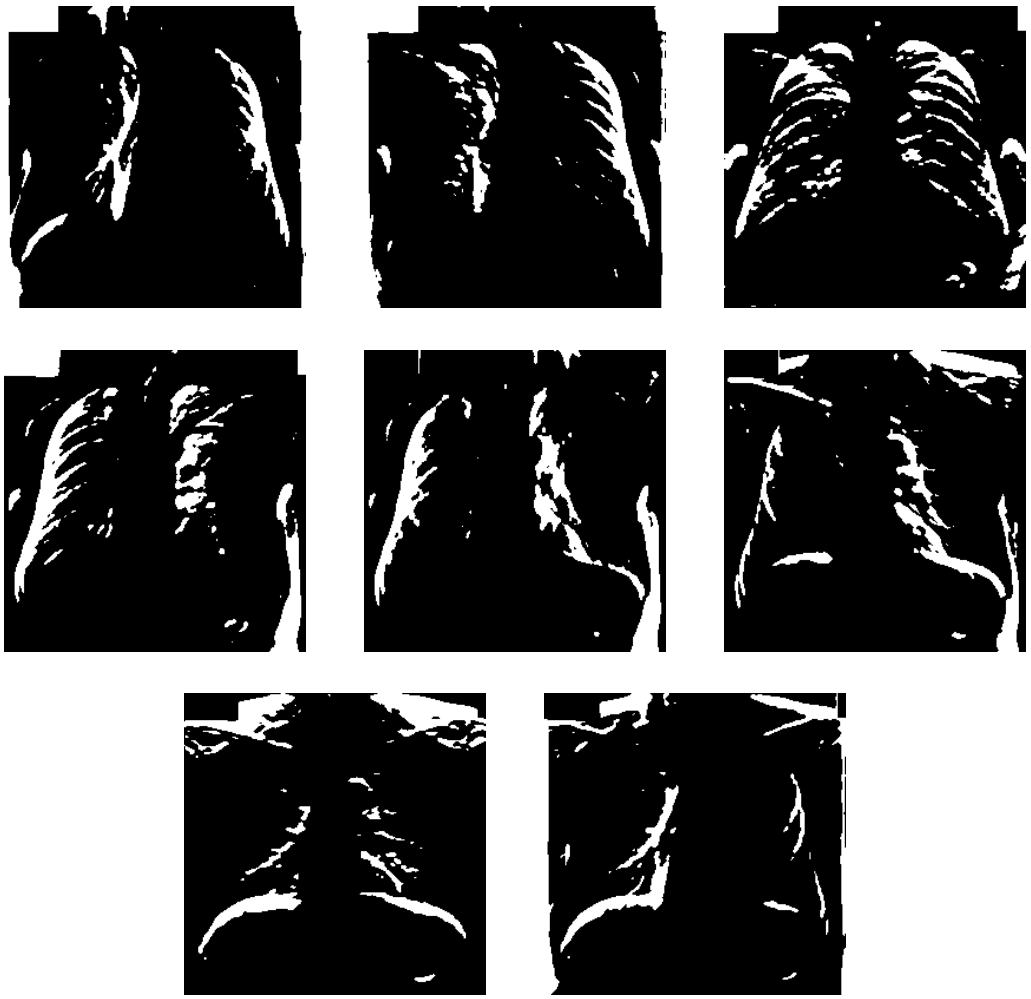

Figure A.5: *Binary results after the application of the oriented Gaussian Filters*

Appendix B

Laplacian of Gaussian Operators

Definition
The Laplacian of a 2D function $f(x, y)$ is a second order derivative defined as:

$$\nabla^2 f(x,y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} \tag{B.1}$$

Eq. B.1 may be implemented in digital form in various ways. The horizontal derivative can be approximated by:

$$\frac{\partial f(x,y)}{\partial x} \approx \Delta_x f(x,y) = f(x,y) - f(x-1,y)$$

and the horizontal second order derivative by:

$$\begin{aligned}
\frac{\partial^2 f(x,y)}{\partial x^2} &\approx \Delta_{xx} f(x,y) \\
&= \Delta_x f(x,y) - \Delta_x f(x+1,y) \\
&= f(x,y) - f(x-1,y) - f(x+1,y) + f(x,y) \\
&= 2f(x,y) - f(x-1,y) - f(x+1,y)
\end{aligned}$$

The vertical derivative can be approximated by:

$$\frac{\partial f(x,y)}{\partial y} \approx \Delta_y f(x,y) = f(x,y) - f(x,y-1)$$

and the vertical second order derivative by:

$$\begin{aligned}
\frac{\partial^2 f(x,y)}{\partial y^2} &\approx \Delta_{yy} f(x,y) \\
&= \Delta_y f(x,y) - \Delta_y f(x,y-1) \\
&= f(x,y) - f(x,y-1) - f(x,y+1) + f(x,y) \\
&= 2f(x,y) - f(x,y-1) - f(x,y+1)
\end{aligned}$$

from the last equations the following formulation for the Laplacian operator can be derived:

$$\nabla^2 f(x,y) \approx 4f(x,y) - [f(x+1,y) + f(x-1,y) + f(x,y-1) + f(x,y+1)]$$

this corresponds to the convolution of $f(x,y)$ with the filter with the mask LP:

$$\nabla^2 f(x,y) = f(x,y) * LP$$

where

$$LP = \begin{array}{|c|c|c|}\hline 0 & -1 & 0 \\ \hline -1 & 4 & -1 \\ \hline 0 & -1 & 0 \\ \hline \end{array}$$

Since the Laplacian is a derivative operator, the sum of the coefficients has to be zero. Hence the response is zero whenever the point in question and its neighbors have the same value.

A more general use of the Laplacian is in finding the *location* of edges using its zero-crossing property. This concept is based on convolving an image with the laplacian of a 2D Gaussian function of the form:

$$g_\sigma(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2+y^2/2\sigma^2)} \tag{B.2}$$

Let $r^2 = x^2 + y^2$, then the gaussian takes the form:

$$g_\sigma(r,\vartheta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-r^2/2\sigma^2)} \tag{B.3}$$

where σ is standard deviation. Then from Eq. B.1 the Laplacian of g (that is, the second derivative of g with respect to r) is

$$\nabla^2 g(r,\vartheta) = \frac{1}{\sqrt{2\pi\sigma^2}} (\frac{r^2-\sigma^2}{\sigma^4}) e^{-(r^2/2\sigma^2)} \tag{B.4}$$

Because of its shape, the laplacian of gaussian is called *mexican hat* and it is shown in Figure B.1.

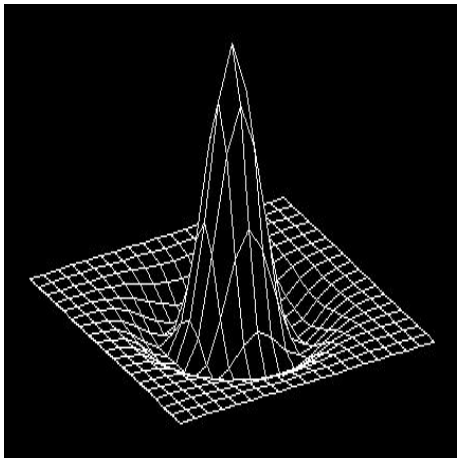

Figure B.1: *The mexican Hat function*

Appendix C

Sensitivity, Specificity and ROC analysis

Considering a *confusion matrix* for a dichotomic problem, the test results can be subdivided in 4 categories as shown in table C.1. The columns refer to the true (expected) values and the rows to the predicted values. P stands for *Positive* examples and N stands for *Negative* examples. True Positives (TP) are Positive examples correctly classified as Positive; True Negatives (TN) are Negative examples correctly classified as Negative; False Positives (FP) are Negative examples incorrectly classified as Positive, and False Negatives (FN) are Positives incorrectly classified as Negative. Using these distinct notions of correct and incorrect classification, we can define different quantities to evaluate the performance of a classifier.

Table C.1: Confusion matrix for a dichotomic problem.

		Expected	
		P	N
Predicted	P	TP	FP
	N	FN	TN

The *Sensitivity* expresses the ratio between the correctly predicted Positive examples and the total number of the Positive examples:

$$Sensitivity = \frac{TP}{TP + FN} \tag{C.1}$$

In the literature this quantity is also called *Recall*.
The *Specificity* expresses the ratio between the correctly predicted Negative examples and the total number of the Negative examples:

$$Specificity = \frac{TN}{TN + FP} \tag{C.2}$$

The *Precision* expresses the ratio between the correctly predicted Positive examples and the total number of examples predicted as Positive:

$$Precision = \frac{TP}{TP + FP} \tag{C.3}$$

The complementary of the Specificity (*1 - Specificity*) is the ratio between the examples incorrectly predicted as Positive and the total number of Negative examples, i.e. it expresses the fraction of the incorrectly classified Negative examples with respect to the total number of Negative examples:

$$1 - Specificity = 1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP} \tag{C.4}$$

Using the above notation, the *Accuracy* is the ratio between the number of correctly classified examples and the total number of examples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{C.5}$$

C.0.1 Basic concepts of ROC curve

The *Receiver Operating Characteristics* (ROC) curve has been introduced by the signal processing community in order to evaluate the capability of a human operator to distinguish informative radar signals from noise [22]. At the present, it is mostly used in the medical decision making community for assessing the usefulness of a diagnostic test.
In order to express in a synthetic way the performance of a classifier system, the Receiver Operating Characteristic (ROC) analysis offers a suitable tool to jointly evaluate sensitivity and specificity: it can be understood as a plot of the probability of classifying correctly the Positive examples against the rate of incorrectly classifying True negative examples. In this sense, one can interpret this curve as a comparison of the classifier across the entire range of class distributions and error costs. In ROC analysis the performance of a classifier is defined through pairs of Sensitivity and 1-Specificity values. Hence, in this two-dimensional ROC space the performance of a certain classifier is defined by a point, i.e. by its 1-Specificity (X-axis) and Sensitivity (y-axis). In the case of classifiers obtained by thresholding, such as MLP or SVM, the *ROC curve* can be computed by varying the decision threshold of the classifier, which describes the trade-off between Specificity and Sensitivity.
Using ROC curves the performance of different learning systems can be compared: the best point in the ROC plane is $(0, 1)$, i.e. $1 - Specificity = 0$ and $Sensitivity = 1$; the worst point is the opposite $(1, 0)$; ROC curves lying near the diagonal correspond to random guessing classifiers, and in general learning systems with ROC curves lying on the the top and leftmost portion of the ROC plane are the better ones. Figure (C.1) depicts an example of

the ROC curve of a given classifier.
The most frequently used performance measure extracted from the ROC curve is the value of the *Area Under the Curve*, commonly denoted as *AUC*. When AUC is equal to 1, the classifier achieves perfect accuracy if the threshold is correctly chosen, and a classifier that predicts the class at random has an associated AUC of 0.5.

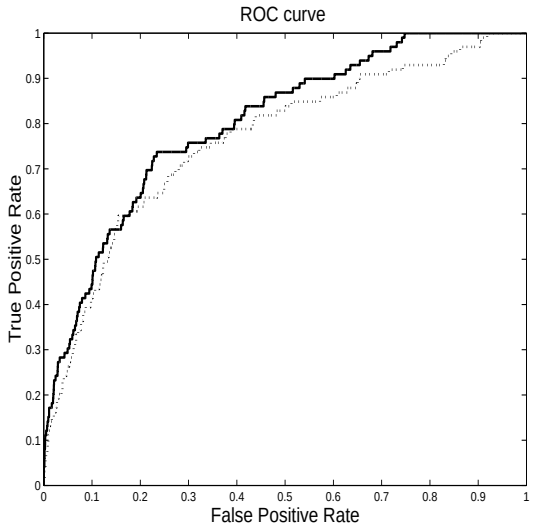

Figure C.1: *Example of ROC obtained with 2-norm SVMs. the solid curve correspond to a gaussian kernel and the dotted one is the ROC curve of a polynomial kernel.*

Appendix D

Support Vector Machines

In this section we provide a very brief overview of Support Vector Machines in order to introduce the main notions and concepts used in the thesis. See [88] and [18] for more detailed descriptions.
Given a data set $\mathcal{Z} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, $\mathbf{x}_i \in \mathbb{R}^N, y_i \in \mathcal{Y} = \{-1, 1\}$, where y_i are the labels of two different classes of examples, a linear classifier computes a decision function $g(\mathbf{x}) = \text{sign}(\text{f}(\mathbf{x}))$, where $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$.

For a point $\mathbf{x}_p$ on the separating hyperplane $f(\mathbf{x}_p) = \mathbf{w} \cdot \mathbf{x}_p + b = 0$ (See

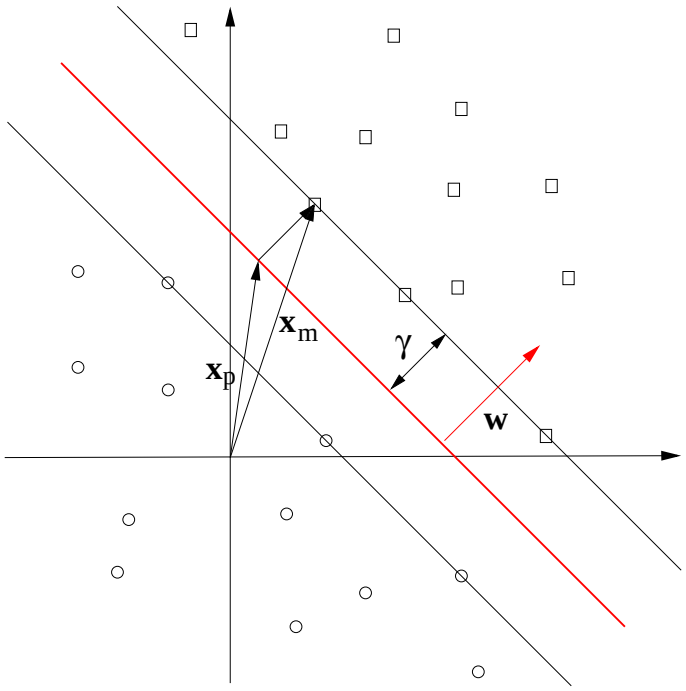

Figure D.1: *Separating hyperplane and margins in a two-class classification problem*

Figure D.1), a point $\mathbf{x}_m$ on the margin whose width is γ can be expressed as:

$$\mathbf{x}_m = \mathbf{x}_p + \frac{\mathbf{w}}{||\mathbf{w}||}\gamma$$

Then $f(\mathbf{x}_m) = \mathbf{w} \cdot \mathbf{x}_m + b = \mathbf{w} \cdot \mathbf{x}_p + \frac{\mathbf{w} \cdot \mathbf{w}}{||\mathbf{w}||}\gamma + b = \gamma||\mathbf{w}||$. The *functional margin* is $\gamma||\mathbf{w}||$ and the *geometric margin* is $\gamma = \frac{f(\mathbf{x}_m)}{||\mathbf{w}||}$.
To obtain the *canonical separating hyperplane* we need to normalize w.r.t the functional margin:

$$f_c(\mathbf{x}) = \frac{f(\mathbf{x})}{\gamma||\mathbf{w}||}$$

The *canonical functional margin* is

$$f_c(\mathbf{x}_m) = \frac{f(\mathbf{x}_m)}{\gamma||\mathbf{w}||} = 1$$

The *canonical margin* is $\gamma_c = \frac{1}{||\mathbf{w}||}$.
From this point we consider only the canonical hyperplane (that is the hyperplane with canonical margin $1/||\mathbf{w}||$.
In order to maximize the margin $\gamma = \frac{1}{||\mathbf{w}||}$ and to correctly separate the examples we need to solve a constrained quadratic optimization problem:

$$\begin{array}{ll} \text{Minimize} & \mathbf{w} \cdot \mathbf{w} \\ \text{subject to} & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \\ & 1 \leq i \leq n \end{array}$$

The hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ that solves this quadratic optimization problem is the *maximal margin hyperplane* with margin $\gamma = \frac{1}{||\mathbf{w}||}$.
The lagrangian associated with the primal optimization problem is:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w} \cdot \mathbf{w} - \sum_{i=1}^{n} \alpha_i (y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1)$$

leading to this set of optimality conditions:

$$\begin{aligned} \frac{\partial L(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i = \mathbf{0} \\ \frac{\partial L(\mathbf{w}, b, \alpha)}{\partial b} &= \sum_{i=1}^{n} y_i \alpha_i = 0 \end{aligned}$$

hence

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i \\ 0 &= \sum_{i=1}^{n} y_i \alpha_i \end{aligned}$$

Putting the relations obtained into the primal we have:

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2}\mathbf{w}\cdot\mathbf{w} - \sum_{i=1}^{n}\alpha_i(y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^{n}\alpha_i \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) \end{aligned}$$

obtaining the associated dual optimization problem:

$$\begin{array}{ll} \text{Maximize} & \Phi(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\ \text{subject to} & \sum_{i=1}^{n} y_i \alpha_i = 0 \\ & \alpha_i \geq 0, \quad 1 \leq i \leq n \end{array}$$

The hyperplane whose weight vector $\mathbf{w}^* = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i$ solves this quadratic optimization problem is the *maximal margin hyperplane* with geometric margin $\gamma = \frac{1}{||\mathbf{w}||}$.
The linear SVMs compute the family of linear functions:

$$\mathcal{F}(\mathbf{x}, \mathbf{w}, b) = \{\mathbf{x}\cdot\mathbf{w} + b, \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$$

If α^* is the solution of the dual optimization problem then

- $\mathbf{w}^* = \sum_{i=1}^{n} y_i \alpha_i^* \mathbf{x}_i$ is the weight vector of the maximal margin hyperplane
- $f(\mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x} + b^* = \sum_{i=1}^{n} y_i \alpha_i^* \mathbf{x}_i \cdot \mathbf{x} + b^*$ is the corresponding discriminant function.
- The decision function $g : \mathbb{R}^n \to \{-1, +1\}$ is $g(\mathbf{x}) = \text{sign}(\sum_{i=1}^{n} y_i \alpha_i^* \mathbf{x}_i \cdot \mathbf{x} + b^*)$

The SVM algorithm minimizes both the empirical risk and the confidence interval [88]. Indeed, maximizing the margin, that is equivalently minimizing $||\mathbf{w}||$, we minimize the Vapnik Chervonenkis (VC) dimension, and the confidence interval depends mainly on the ratio (VC) dimension/cardinality of the training set. In order to consider non linearly separable data we need to introduce soft margin SVM and kernels. In this setting we first add to the primal optimization problems a set of slack variables ξ_i, and a

$$\begin{array}{ll} \text{Minimize} & \mathbf{w}\cdot\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \\ & 1 \leq i \leq n \end{array}$$

If $K(\mathbf{x}, \mathbf{x}')$ is a symmetric function satisfying *Mercer's conditions*, that is:

$$\int \int K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0$$

for all f such that $\int f^2(\mathbf{x}) d\mathbf{x} < \infty$, then we can expand $K(\mathbf{x}, \mathbf{x}')$ in a some inner product feature space:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j \phi(\mathbf{x}) \phi(\mathbf{x}')$$

Note that in the dual representation of linear SVMs the inputs appears only in a dot-product form:as a consequence we can substitute the dot-products in the input space with a kernel function obeying Mercer's conditions:

$$\begin{array}{ll} \text{Maximize} & \Phi(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i \mathbf{x}_j) \\ \text{subject to} & \sum_{i=1}^{n} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad 1 \leq i \leq n \end{array}$$

The discriminant function obtained from the solution of this quadratic optimization problem is:

$$f(\mathbf{x}, \alpha^*, b) = \sum_{i=1}^{n} y_i \alpha_i^* K(\mathbf{x}_i, \mathbf{x}) + b^*$$

The SVM receives as inputs patterns $\mathbf{x}$ in the input space, but works in a high dimensional (possibly infinite) feature space, where it performs a linear separation of the data. The symmetric function $K(\cdot, \cdot)$ must be chosen among the kernels of Reproducing Kernel Hilbert Spaces [93]; three possible choices are:

- Linear kernel: $K(\mathbf{u}, \mathbf{v}) = \mathbf{u} \cdot \mathbf{v}$
- Polynomial kernel: $K(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \cdot \mathbf{v} + 1)^d$
- Gaussian kernel: $K(\mathbf{u}, \mathbf{v}) = \exp(-\|\mathbf{u} - \mathbf{v}\|^2 / \sigma^2)$

The bias and variance of SVMs are typically controlled by two parameters. The parameter C controls the tradeoff between fitting the data (achieved by driving the ξ_i's to zero) and maximizing the margin (achieved by driving $\|\mathbf{w}\|$ to zero). Setting C large should tend to minimize bias. The second parameter that controls bias arises only in SVMs that employ parameterized kernels such as the polynomial kernel (where the parameter is the degree d of the polynomial) and RBF kernels (where the parameter is the width σ of the gaussian kernel). Bias and variance depend critically on these parameters [83].

Appendix E

Gabor Filters

Simple cells in the primary visual cortex have receptive fields (RFs) which are restricted to small regions of space and are highly structured (see the work by Marcelja described in [55]). Earlier examinations by Hubel and Wiesel leaded to a description of these cells as edge detectors. More recent examinations, among others the one by Jones and Palmer in 1987 (see [37]), showed that the response behaviour of simple cells of cats corresponds to local measurements of frequencies. Furthermore Pollen and Ronner (in [66]) examined the phase relation of adjacent cells in the visual cortex of cats. They concluded that that most of the simple cells can be combined in pairs, one cell of each pair has an RF modelled by an even function and the other one can be modelled with an odd function. This allows to model both RFs of such a pair of cells by a complex-valued function; the authors suggested the use of a Cosine function and a Sine function. Indeed, if the modelled RFs of both cells are combined in a complex notation with

$$exp(ikx) = cos(kx) + isin(kx) \tag{E.1}$$

the real part corresponds to the cell with even symmetry and the imaginary part to the cell with odd symmetry. This observation leaded to the formulation of a biologically motivated filter, approximating the characteristics of the RFs of the cells in the visual cortex, and called the Gabor filter (see [19]). A Gabor filter can be viewed as a sinusoidal plane of particular frequency and orientation, modulated by a gaussian envelope. The sinusoidal plane is expressed by a complex sinusoid, known as a carrier, meanwhile the gaussian shaped function is commonly referred as the envelop. The mathematical formulation of the $2-D$ Gabor filter is:

$$\psi_j(\vec{x}) = \frac{k_j^2}{\sigma^2} exp\left(-\frac{k_j^2 \left|\vec{x}\right|^2}{2\sigma^2}\right)\left[exp(i\vec{k_j}\vec{x}) - exp(-\frac{\sigma^2}{2})\right] \tag{E.2}$$

where $\vec{x} = (x, y)$ and the vector:

$$\vec{k_j} = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_\nu \cos\varphi_\mu \\ k_\nu \sin\varphi_\mu \end{pmatrix}$$

determines the scale (k_ν) and the orientations ($\cos\varphi_\mu$ and $\sin\varphi_\mu$).
The term $\frac{k_j^2}{\sigma^2}$ compensates for the frequency-dependent decrease of the power spectrum in natural images (see [69]). The term $exp\left(-\frac{k_j^2|\vec{x}|^2}{2\sigma^2}\right)$ is the Gaussian envelope function restricting the sinusoidal plane. The first term in the parenthesis of equation E.2 is the complex-valued sinusoidal plane itself; note that the second term in parenthesis makes the function DC-free (i.e with a zero mean):

$$\int \psi_j(\vec{x}) dx^2 = 0$$

This property ensures the robustness of the Gabor filters with respect to the variations of the illumination.
In addition the Gabor filters have received considerable attention since they have been shown to have good localization properties in both spatial and frequency domain and thus are well suited for texture segmentation problems ([34], [35]). Gabor filters have been used in many applications, such as texture segmentation, target detection, fractal dimension management, document analysis, edge detection, retina identification, image coding and image representation. ([94]).

E.0.2 Gabor filter bank

We realized a Gabor filter bank at 5 different scales and 8 orientations by setting in $\vec{k_j}$ the following values for the parameters:

$$k_\nu = 2^{-\frac{\nu+2}{2}}\pi, \quad \nu = 0, 1, 2, 3, 4$$

$$\varphi_\mu = \mu\frac{\pi}{8}, \quad \mu = 0, 1, 2, 3, 4, 5, 6, 7$$

The subscript index j, in equation E.2, is then defined as $j = \mu + 8\nu$; the width of the gaussian function used is σ/k_ν, where we set $\sigma = 2\pi$.

Figure E.1 shows the real parts of the Gabor filters obtained by setting a certain orientation value and varying the scale in the range established; Figure E.2 shows the 8 orientations of the real parts of the Gabor filters when the same scale is kept constant. Figures E.3 and E.4 represents in the same order the imaginary parts of the same Gabor filters.
Note that the real part of the Gabor filter is always an even function, meanwhile the imaginary part is always odd.

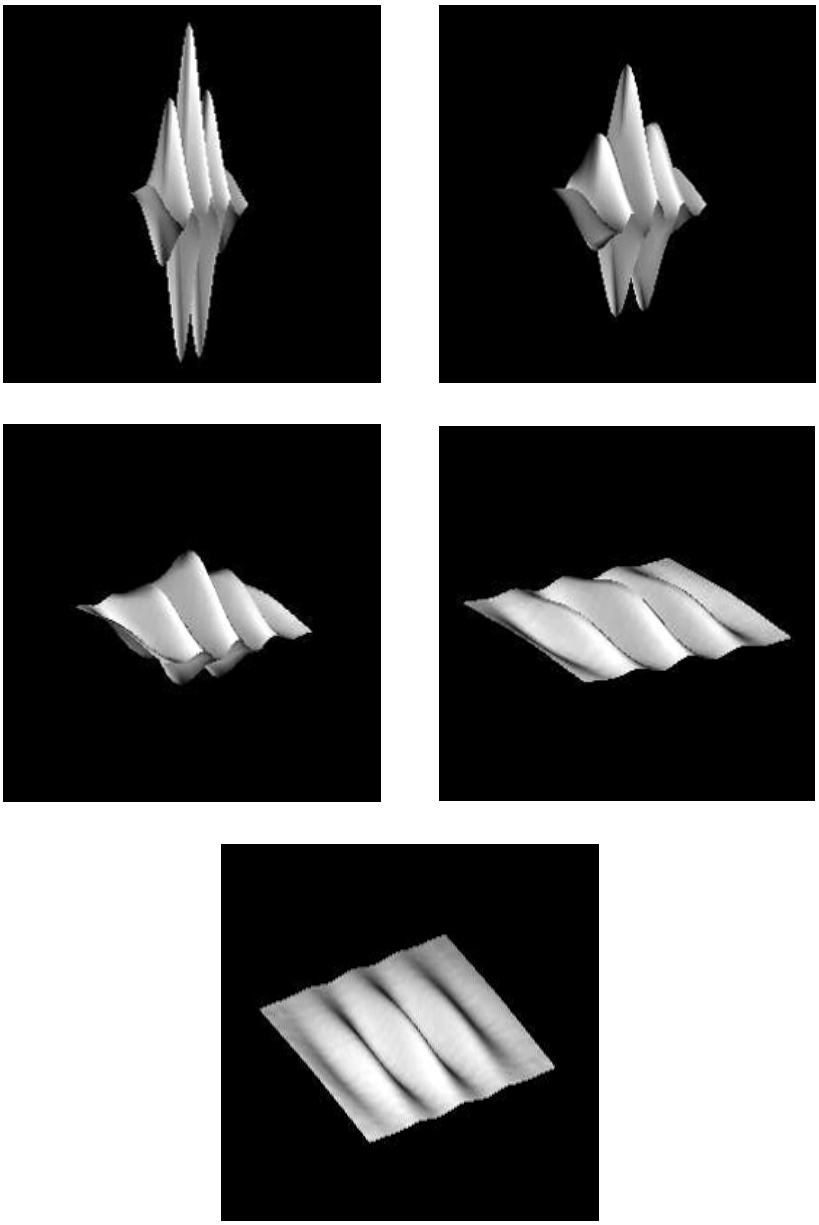

Figure E.1: Real part of the Gabor filter at 5 different scales

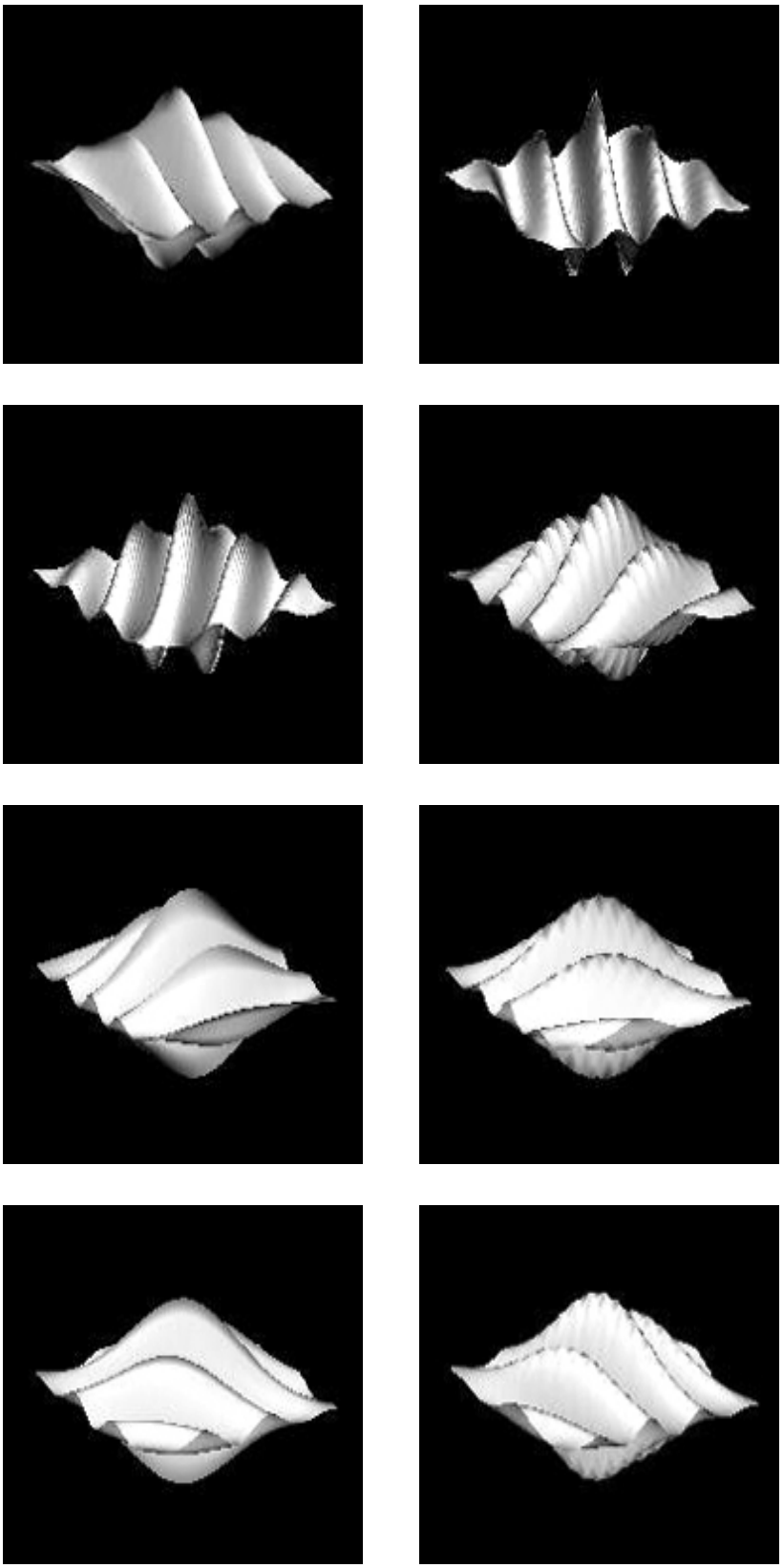

Figure E.2: Real part of the Gabor filters in 8 different orientations

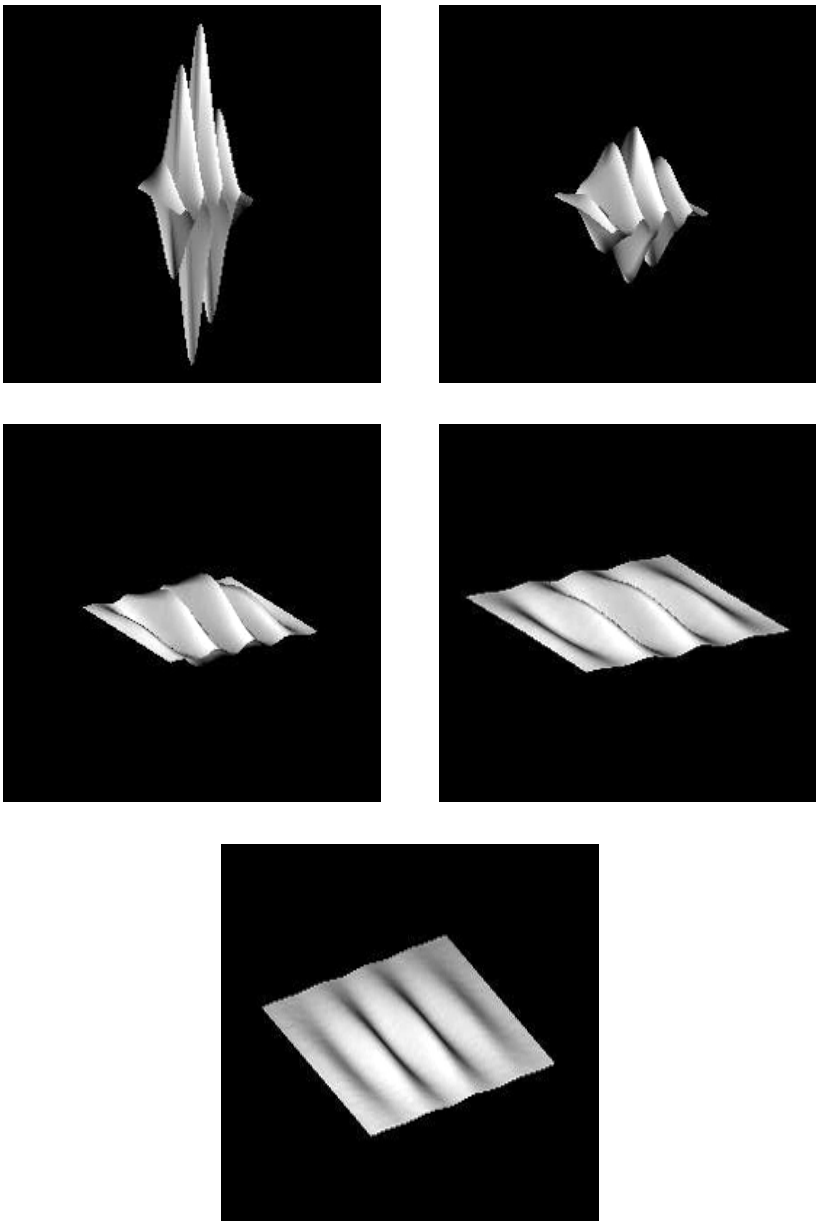

Figure E.3: Imaginary part of the Gabor filter at 5 different scales

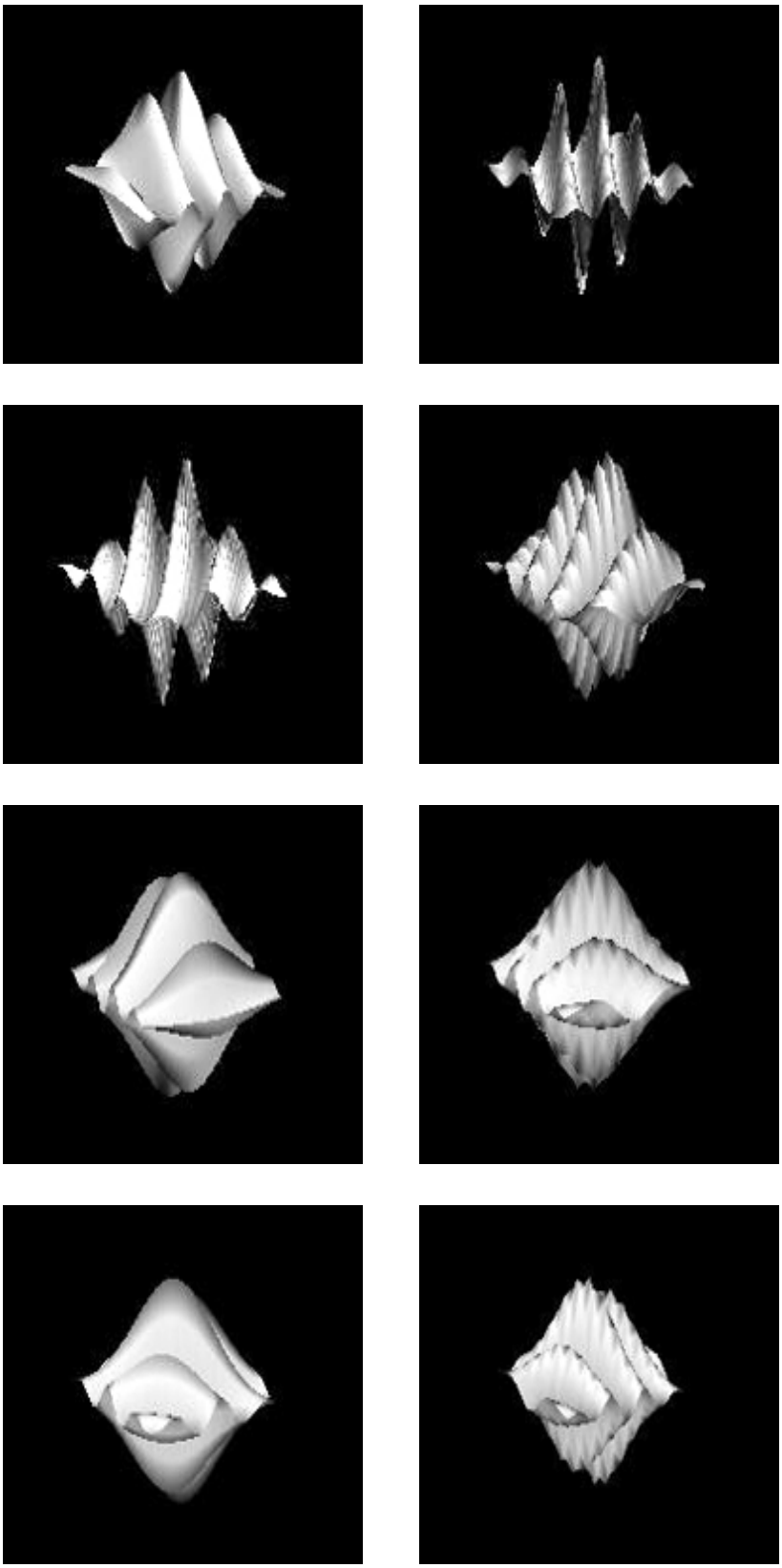

Figure E.4: Imaginary part of the Gabor filters in 8 different orientations

Bibliography

[1] B.Van Ginneken A. Schilham and M. Loog. Multi-scale nodule detection in chest radiographs. *Proc. MICCAI*, 2003.

[2] M.L: Abercrombie. *The anatomy of judgment.* Hutchinsons, LONDON, 1960.

[3] S.G. Armato, M.Giger, K. Ashizawa, and H.MacMahon. Automated lung segmentation in digitital lateral chest radiographs. *Medical Physics*, 24:1507–1520, 1998.

[4] S.G. Armato, M.Giger, and H.MacMahon. Automated lung segmentation in digitized posteroanterior chest radiographs. *Academic radiology*, 5:245–255, 1998.

[5] Y. Asai, Y.Tanabe, Y. Ozaki, H. Kubota, M. Matsumoto, and H. KanaMori. Optimum tube voltage for chest radiographs for chest radiographs obtained by psycophysical analysis. *Medical Physics*, 25:2170–2175, 1998.

[6] J.H.M. Austin, B.M. Romeny, and L.S. Goldsmith. Missed bronchogenic carcinoma: radiographic findings in 27 patients with apotentially resectable lesion evident in retrospect. *Radiology*, 182:115–122, 1992.

[7] H.C. Becker, W.J. Nettleton, P.H. Meyers, and Jr. C.M.Nice. Digital computer determination of a medical diagnostic index directly from chest x-ray images. *IEEE Transaction on Biomedical Engineering*, BME-11:67–72, 1964.

[8] Keserci Bilgin and Yoshida Hiroyuki. Computerized detection of pulmonary nodules in chest radiographs based on morphological features and wavelet snake model. *Medical Image Analisys*, 6:431–447, 2002.

[9] M.S. Brown, L.S. Wilson, B.D. Doust, R.W. Gill, and C.Sun. Knowledge-based method for segmentation and analysis of lung boundaries in chest x-rays images. *Computerized Medical Imaging and Graphics*, 22:463–477, 1998.

[10] P.E. Buell. The importance of tumor size in prognosis for resected bronchogenic carcinoma. *Journal of Oncological Surgery*, 3:539–551, 1971.

[11] R.E. Bung and C.L. Herman. Use of diagnostic imaging procedures: a nationwide hospital study. *Radiology*, 163:569–573, 1987.

[12] Paola Campadelli and Elena Casiraghi. Lung edge detection in postero anterior chest radiographs. *Proc. of Image: e-learning, understanding, information retrieval, medical*, 15:27–37, 2003.

[13] J. Canny. A computational aprroach to edge detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 8:679–698, 1986.

[14] F.M. Carrascal, J.M. Carreira, M. Souto, P.G. Tahoces, L. Gomez, and J.J. Vidal. Automatic calculation of total lung capacity from automatically traced lung boundaries in postero-anterior and lateral digital chest radiographs. *Medical Physics*, 25:1118–1131, 1998.

[15] D. Cheng and M. Goldberg. An algorithm for segmenting chest radiographs. *Proc SPIE*, pages 261–268, 1988.

[16] R.W. Conners, C.A. HArlow, and S.J. Dweyer III. Radiographic image analysis: past and present. *Radiology*, 1982.

[17] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Comput. Vis, Image Understanding*, 61:38–59, 1995.

[18] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK, 2000.

[19] J.D. Daugman. Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 36:1169–1179, 1988.

[20] J.S. Duncan and N. Ayache. Medical image anlysis: progress over two decades and the challenges ahead. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22:85–106, 2000.

[21] J. Duryea and J.M. Boone. A fully automatic algorithmfor the segmentation of lung fields in digital chest radiographic images. *Medical Physics*, 22:183–191, 1995.

[22] J. Egan. *Signal Decision Theory and ROC Analysis*. Academic Press, 1975.

[23] J. Forrest and P. Friedman. Radiologic errors in patient with lung cancer. *West Journal on Med.*, 134:485–490, 1981.

[24] W.T. Freeman and E.H.Adelson. The design and use of steerable filters. *IEEE Trans. on PAttern Analysis and Machine Intelligence*, 13:891–906, 1991.

[25] G. Gavelli and E. Giamplama. Sensitivity and specificity of chest x-ray screening for lung cancer. *Proceedings of the international conference on Prevention and Early Diagnosis of Lung Cancer*, Varese (ITALY):103–108, 1998.

[26] M. Giger, K. Doi, and H. Mac Mahon. Image feature analisys and computer-aided diagnosis in digital radiography:automated detection of nodules in peripheral lung fields. *Med. Phisycs*, 15:158–166, 1988.

[27] M. Giger, K. Doi, H. Mac Mahon, C. Metz, and F.Yin. Pulmonary nodules: Computer-aided detection in digital chest images. *Radiographics*, 10:41–51, 1990.

[28] M. L. Giger, N. Ahn, K. Doi, H. MacMahon, and C. E. Metz. Computerized detection of pulmonary nodules in digital chest images: use of morphological filters in reducing false positive detections. *Med. Phys.*, 17:861–865, 1990.

[29] L. W. Guiss and P. Kuenstler. A retrospective view of survey photofluorocarbons of persons with lung cancer. *Cancer*, 13:91–95, 1960.

[30] J.W. Gurney and S.J.Swensen. Solitary pulmonary nodules: determining the likelihood of malignancy with neural network analysis. *Radiology*, 196:823–829, 1995.

[31] R.M. Haralick, K. Shanmugan, and I. Dinstein. Textural features for image classification. *IEEE Transaction on systems, Man and Cybernetics*, 3:610–621, 1973.

[32] A. Hasegawa, S.-C. Lo, M.T. Freedman, and S.K. Mun. Convolution neural network based detection of lung structure. *Proc. SPIE 2167*, pages 654–662, 1994.

[33] T. Heelan, B.J. Flehinger, M.R. Melamed, M.B. Zaman, W.B. Perchick, J.F. Caravelli, and N. Martini. Non-small-cell lung cancer: result of the new york screening program. *Radiology*, 151:289–293, 1984.

[34] A. Jain and S. Bhattacharjee. Address block location on envelopes using gabor filters. *Signal Processing*, 25, 1992.

[35] A. Jain, N. Ratha, and S. Lakshmanan. Object detection using gabor filters. *Pattern Recognition*, 30:295–309, 1997.

[36] A. K. Jain, Y. Zhong, and M.P. Dubuisson-Jolly. Deformable template models: a review. *Signal Processing*, 71:109–129, 1998.

[37] Jones and Palmers. An evaluation of the two dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiologic*, 58:1233–1258, 1987.

[38] D.Catarious Jr., A.Baydush, and Jr. C.Floyd. Initial development of a computer aided diagnosis tool for solitary polmunary nodules. *Proc SPIE*, 4322:710–717, 1988.

[39] J.Wei, Y. Hagihara, and H. Kobatake. Detection of cancerous tumors on chest x-ray images- candidates detection filtering and its evaluation. *Presented at the Int. Conf. on Image. Proc (ICIP '99)*, 1999.

[40] J.Wei, Y. Hagihara, A. Shimizu, and H. Kobatake. Optimal image feature set for detecting lung nodules on chest x-rays images. *CARS 2002*, 2002.

[41] A. Kano, K. Doi, H. MacMahon, D.D. Hassel, and M.Giger. Digital image subtraction of temporally sequential chest images for detection of interval change. *Med. Physics*, 21:453–461, 1994.

[42] J. N. Kapur, P. K. Sahoo, and A.K. C.Woong. A new method for gray level picture thresholding using the entropy of the histogram. *Computer Vision Graphics and Image Processing*, 29:273–285, 1985.

[43] R. Klette and P.Zamperoni. *Handbook of image processing operators*. Wiley, 1994.

[44] K.Morik, P. Brockhausen, and T. Joachims. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. *International Conference on Machine Learning (ICML)*, 1999.

[45] H. Kundel and G. Revesz. *Optimization of Chest RAdiography*. HHS Publication, FDA, RockVille, MD, 1980.

[46] K.D. Laws. Texture image segmentation. *Phd thesis, University of Souther California*, 1980.

[47] J.-S. Lin, S.-C. Lo, M. Freedman, and S.Mun. Reduction of false positives in lung nodule detection using a two-level neural classification. *IEEE Trans. Med. Imag.*, 15:206–217, 1996.

[48] T. Lindeberg. *Scale-Space theory in Computer vision*. The Kluwer International Series in Engineering and Computer Science, The Kluwer Academic Publishers, 1994.

[49] T. Lindeberg. Scale-space: A framework for handling image structures at multiple scales. *Proc. CERN, School of Computing*, 8-21 September, The Nederlands, 1996.

[50] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30, 1998.

[51] S.-C. Lo, M. Freedman, J.-S. Lin, and S.Mun. Automatic lung nodule detection using profile matching and backpropagation neural network techniques. *Journal Digital Imaging*, 1:48–54, 1993.

[52] S.-C. Lo, S.L.Lou, J.-S. Lin, M. Freedman, M.Chien, and S.Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Trans. on Med. Imag.*, 14:711–718, 1995.

[53] H. MacMahon and K. Doi. Digital chest radiography. *Clin. Chest Med.*, 12:19–32, 1991.

[54] H. MacMahon, K. Doi, R. Engelmann, F.M. Behlen, K.R. Hoffman, T. Ishida, C. Roe, and C.E. Metz. Computer-aided diagnosis of pulmonary nodules: results of a large-scale observer study. *Radiology*, 23:723–726, 1999.

[55] S. Marcelja. Mathematical description of the response of simple cortical cells. *Journal of Optical Society of America*, A 70(11):1297–1300, 1980.

[56] T. Matsumoto, H. Yoshimura, K. Doi, M.L. Giger, A. Kano, H. MacMahon, M. Abe, and S. Montner. Image features analysis of false-positives diagnosis produced by automated detection of lung nodules. *Investigative Radiol.*, 27:587–579, 1992.

[57] M.F. McNitt-Gray, H.K. Huang, and J.W. Sayre. Feature selection in the pattern classification problem of digital chest radiographs segmentation. *IEEE Trans. on Med. Imaging*, 14:537–547, 1995.

[58] M.F. McNitt-Gray, J.W. Sayre, H.K. Huang, and M. Razavi. A pattern classification approach to segmentation of chest radiographs. *PROC SPIE 1898*, pages 160–170, 1993.

[59] P.H. Meyers, Jr. C.M. Nice, H.C. Becker, W.J. Nettleton, J.W. Sweeney, and G.R. Meckstroth. Automated computer analysis of radiographic images. *Radiology*, 83:1029–1034, 1964.

[60] J. R. Muhm, W.E. Miller, R.S. Fontana, D.R. Sanderson, and M.A. Uhlenhopp. Lung cancer detected during a screening program using fourth-month chest radiographs. *Radiology*, 19:609–615, 1983.

[61] G.P. Murphy, Jr. W. Lawrence, and Jr. R.E. Lenhard. Clinical oncology. *American Cancer Society*, Washington DC, 1995.

[62] K. Nakamura, H. Yoshida, R. Engelmann, H. MacMahon, S. Katsuragawa, T. Ishida, K. Ahizawa, and K. Doi. Computerized analisys of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks. *Radiology*, 214:823–830, 2000.

[63] M. Penedo, M. Carreira, A. Mosquera, and D. Cabello. Computer aided diagnosis: A neural network based approach to lung nodule detection. *IEEE Trans. Med. Imag.*, 17:872–880, 1998.

[64] M. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12:629–639, 1990.

[65] E. Pietka. Lung segmentation in digital chest radiographs. *Journal of digital imaging*, 7:79–84, 1994.

[66] Daniel A. Pollen and Steven F. Ronner. Phase relationship between adjacent simple cells in the visual cortex. *Science*, 212:1409–1411, 1981.

[67] R.C.Gonzales and R.E.Woods. *Digital Image Processing.* Addison Wesley Publishing Company, USA, 1992.

[68] W.C. Rontgen. Ueber eine neue art on strahlen. *Sitzungsberichte der Physikalisch-Medicinisch Gesellshaft zu Wurzburg*, 15:132–141, 1895.

[69] Daniel L. Ruderman and William Bialek. Statistics of natural images: Scaling in the wood. *Physical Review Letters*, 73(6):814–817, 1994.

[70] E. Samei, M.J.Flynn, and W.Eyler. Detection of subtle lung nodules: relative influence of quantum and anatomic noise on chest radiographs. *Radiology*, 213:727–734, 1999.

[71] E. Samei, M.J.Flynn, W.Eyler, and E. Peterson. Subtle lung nodules: Influence of local anatomic variations on detection. *Radiology*, 228:76–84, 2003.

[72] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi. 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*, 174:71–74, 2000.

[73] M.J. Smith. *Error and Variation in DIagnostic RAdiology.* Thomas Springfield, 1967.

[74] Anna Staffiere. Metodi di segmentazione per l'individuazione dei polmoni in radiografie digitali. *Master Thesis*, Statal University of Milan, Department of Computer Science:Milano, 2003.

[75] J.D. Steele, W.P. Kleitsch, J.E. Dunn, and P. Buell. Survival in males with bronchogenic acrcinamas resected as asymptomatic solitary pulmonary nodules. *Annotation Thoracic surgery*, 2:368–376, 1966.

[76] F. P. Stitik. *Screening for cancer: Chest Radiology.* Miller AB, edition, 1985.

[77] H. Suzuki and N. Inakoa. Development of a computer aided detection system for lung cancer diagnosis. *Proc. SPIE*, 1652:567–571, 1992.

[78] H. Suzuki, N. Inakoa, H. Takabatake, M. Mori, H. Natori, and A. Suzuki. An experimental system for detecting lung nodules by chest x-ray image processing. *Proc. SPIE*, 1450:99–107, 1991.

[79] T.Kobayashi, X.-W. Xu, H. MacMahon, C. Metz, and K. Doi. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. *Radiology*, 199:843–848, 1996.

[80] T.Matsumoto, H.Yoshimura, K.Doi, M.L.Giger, A.Kano, H.MacMahon, M.Abe, and S.Montner. Potential usefulness of computerized nodule detection in screening programs for lung cancer. *Investigative Radiol.*, 27:471–475, 1992.

[81] O. Tsuji, M.T. Freedman, and S.K. Mun. Automated segmentation of anatomic regions in chest radiographs using an adaptive-sized hybrid neural network. *Med. Phys.*, 25:998–1007, 1998.

[82] T. Uchiyama and M. A.Arbib. Color image segmentation using competitive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:1197–1206, 1994.

[83] G. Valentini and T.G. Dietterich. Bias–variance analysis of support vector machines for the development of svm-based ensemble methods. *Journal of Machine Learning Research*, 5:725–775, 2004.

[84] B. van Ginneken. Computer-aided diagnosis in chest radiographs. *P.h.D. dissertation, Utrecht Univ., Utrecht, The Nederlands*, 2001.

[85] B. van Ginneken and B.M. ter H. Romeny. Automatic segmentation of lung fields in chest radiographs. *Medical Physics*, 27:2445–2455, 2000.

[86] B. van Ginneken, B.M. ter H. Romeny, and M.A. Viergever. Computer-aided diagnosis in chest radiography: A survey. *IEEE Trans. On Med. Imag.*, 20:1228–1241, 2001.

[87] J.W. Vance, C.A. Good, C.H. Hodgson, J.W. Kirklin, and R.P. Gage. The solitary pulmonary lesion due to bronchogenic carcinoma: a 3-year follow-up study on 94 surgically treated patients. *Disease Chest*, 36:231–237, 1959.

[88] V. Vapnik. *Statistical Learning Theory.* Wiley Interscience, 1998.

[89] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans. on Pattern Analysis and Machine intelligence*, 13:583–598, 1991.

[90] N.F. Vittitoe, R.Vargas-Voracek, and C.E. Floyd Jr. Identification of lung regions in chest radiographs using markov random field modeling. *Med. Phys.*, 25:976–985, 1998.

[91] N.F. Vittitoe, R. Vargas-Voracek, and C.E. Floyd Jr. Markov random field modeling in posteroanterior chest radiograph segmentation. *Med. Phys.*, 26:1670–1677, 1999.

[92] Cj Vyborny. The aapm/rsna physics tutorial for residents: Image quality and the clinical radiographic examination. *Radiographics*, 17:479–498, 1997.

[93] G. Wahba. *Spline models for observational data.* SIAM, Philadelphia, USA, 1990.

[94] T. P. Weldon, W. E. Higgins, and D. F. Dunn. Gabor filter design for multiple texture segmentation. *Optical Engineering*, 35:2852–2863, 1996.

[95] Y. Wu, K. Doi, M. Giger, C. Metz, and W. Zhang. Reduction of false positives in computerized detection of lung nodules in chest radiographs using convolution neural network techniques and application for lung nodule detection. *J. Digital Imaging*, 7:196–207, 1994.

[96] X.-W. Xu and K. Doi. Image feature anlysis for computer aided diagnosis: accurate determination of ribcage boundaries chest radiographs. *Medical Physics*, 22:617–626, 1995.

[97] X.-W. Xu and K. Doi. Image feature anlysis for computer aided diagnosis: accurate determination of right and left hemidiaphragm edges and delineation of lung field in chest radiographs. *Medical Physics*, 23:1616–1624, 1996.

[98] X.-W. Xu, S. Katsuragawa, A. Ashizawa, H. MacMahon, and K. Doi. Analisys of image features of histograms of edge gradient for false positive reduction in lung nodule detection in chest radiographs. *Proc SPIE*, 3338:318–326, 1998.

[99] H. Yoshida, X. Xu, K. Doi, and M. Giger. Computer-aided diagnosis (cad) scheme for detecting pulmonary nodules using wavelet transforms. *Proc. SPIE*, 2434:621–626, 1995.

[100] H. Yoshimura, M. Giger, K.Doi, H. MacMahon, and S. Montner. Computerized scheme for the detection of pulmonary nodules: A non linear filtering technique. *Investigative radiol.*, 27:124–127, 1992.

[101] H. Zhao, S.C.B. Lo, M.T. Freedman, and Y. Wang. Enhanced lung cancer detection in temporal subtraction chest radiography using directional filtering techniques. *Proc. SPIE*, 4684, 2002.