Lung Nodule Detection in Chest Radiographs

Rui Teixeira Ribeiro

July 2008

Faculdade de Engenharia da Universidade do Porto

Lung Nodule Detection in Chest Radiographs

Rui Agostinho Fernandes Teixeira Ribeiro

M.SC. THESIS

Tese elaborada no âmbito do
Mestrado Integrado em Engenharia Electrotécnica e de Computadores
(Ramo de Telecomunicações)

Orientador: Prof. Dr. Aurélio Campilho

Julho de 2008

To *Clara* and *Agostinho* , my parents.

Abstract

In the last decades, several research projects have been devoted to the development of computer-aided diagnosis (CAD) systems that could help radiologists to detect lung nodules. Despite the great effort, the problem is still open. In this dissertation, a new CAD system is presented for the detection of lung nodules in posterior-anterior chest radiographs.

The lung area considered in this work includes the regions obscured by the mediastinum, the diaphragm and other anatomic structures (hidden regions). Although those regions are usually excluded from the methods presented in the literature, they were included because some lung nodules may be found there. To detect nodules from both visible and hidden regions, we used a CAD system where these regions were assessed separately and the classification results combined in the end.

In sum, after segmenting those lung regions, a multi-scale blob detector was used to nodule candidate detection. At the last stage, the Support Vector Machine (SVM) methodology was employed to candidate classification in the visible region where we obtained a sensitivity of 83% at 6.5 false positives (FP) per image (i.e., per patient) and 90% at 9 FP per image. For the hidden region, a specially conceived classifier – Hyper-polyhedron Classifier – was used, achieving a sensitivity of 82% at 1.8 FP per image.

Resumo

Nas últimas décadas, vários projectos de pesquisa têm-se dedicado ao desenvolvimento de sistemas de diagnóstico auxiliado por computador (*computer-aided diagnosis systems – sistemas CAD*) que poderão ajudar os radiologistas na detecção de nódulos pulmonares. Apesar dos esforços efectuados, o problema continua ainda em aberto. Nesta dissertação, apresenta-se um novo *sistema CAD* para a detecção de nódulos pulmonares em radiografias toráxicas com incidência postero-anterior.

A área do pulmão considerada neste trabalho inclui as regiões ocultas pelo mediastino, pelo diafragma e outras estruturas anatómicas (regiões escondidas). Embora aquelas regiões sejam normalmente excluídas dos métodos descritos em literatura recente, elas foram incluídas uma vez que podem aí encontrar-se alguns nódulos. Para detectar os nódulos tanto na região visível como na escondida, foi usado um *sistema CAD* onde se avaliam separadamente essas regiões. Os resultados da classificação em cada uma das regiões são combinados no final.

Em suma, depois da segmentação da região dos pulmões, um detector de *blobs* ("manchas arredondadas") multi-escala foi usado para a detecção de candidatos a nódulo. Na última fase, foi utilizada a metodologia de *Support Vector Machines (SVMs)* para a classificação dos candidatos na região visível, onde foi obtida uma sensibilidade de 83% com 6.5 falsos positivos (FP) por imagem (i.e., por paciente) e 90% com 9 FP por imagem. Para a região escondida, foi usado um classificador especialmente concebido – Classificador de Hiper-poliedro (*Hyper-polyhedron Classifier*) – tendo-se atingido uma sensibilidade de 82% com 1.8 FP por imagem.

Acknowledgements

I would like to thank Prof. Aurélio Campilho for the several suggestions given during the development of this project and during the revision of this document.

I would also like to thank INEB (Instituto de Engenharia Biomédica - Biomedical Engineering Institute, Porto) for the availability of facilities and the useful interaction with members of the INEB research group: Biomedical Imaging and Vision Computing Group.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Lung cancer is the leading cause for the estimated number of deaths from cancer worldwide. The World Health Organization (WHO) estimated about 1.3 million deaths per year [1] caused only by this cancer. If the disease is detected at an early stage, when it is still localised, the patient's chance of survival will be substantially higher. Thus, early detection of suspicious lesions is important and the posterior-anterior chest radiography can pay an important role. This low-cost technique is still the most common radiological procedure in the world and has the advantage of requiring a fraction of X-ray dose of computed tomography (CT) scans [2]. On the other hand, chest radiographs are also used for a variety of purposes other than lung cancer detection. Because chest radiographs are so widely prescribed, improvements in the detection of lung nodules in chest radiographs could have a significant impact on early lung cancer detection.

However, clinical studies show that radiologists detect pulmonary nodules in radiographs in about only 70-80% of actually positive cases [3]. This fact has motivated the development of computer aided diagnosis (CAD) systems for lung nodule detection.

We should be aware of the difficulties related to this problem. Firstly, there is a wide

range of nodule sizes. In addition, nodules exhibit a large variation in density (visibility on a radiograph) and some nodules are only slightly denser than the surrounding tissue. Lastly, since nodules can appear anywhere in the lung field, they can be obscured by ribs, mediastinum and structures below the diaphragm.

The aim of this dissertation is to develop an approach for a computer aided diagnosis system for detecting lung nodules in posterior-anterior chest radiographs.

The methodology was evaluated in the publicly available JRST (Japanese Society of Radiological Technology) database [4] which contains 247 posterior–anterior chest radiographs (154 are patients with a lung nodule).

1.1 Posterior-Anterior Chest Radiograph

Figure 1.1 shows a typical posterior-anterior chest radiograph (posterior-anterior means that the radiation passes through the patient from back to front). Several anatomic structures are pointed out [5]:

1) The heart;
2) The aortic arch where the aorta bends;
3) The (in this case left) hilum, where the arteries and veins enter the lung;
4) A darker vertical stripe indicates the trachea;
5) Below the lung fields the diaphragm starts;
6) Clavicle;
7) Shoulderblade;
8) Usually stomach gasses 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.

Henceforth, the regions obscured by the mediastinum, the diaphragm and other anatomic structures will be called hidden regions. In figure 1.2 are outlined the visible and the hidden regions.

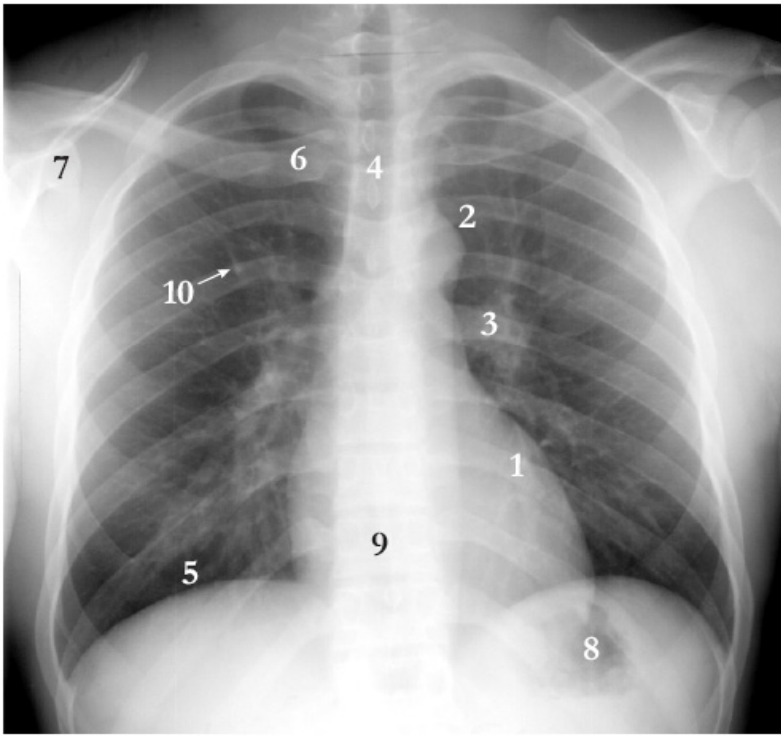


Figure 1.1 - A normal posterior-anterior chest radiograph (this image was taken from [5]).

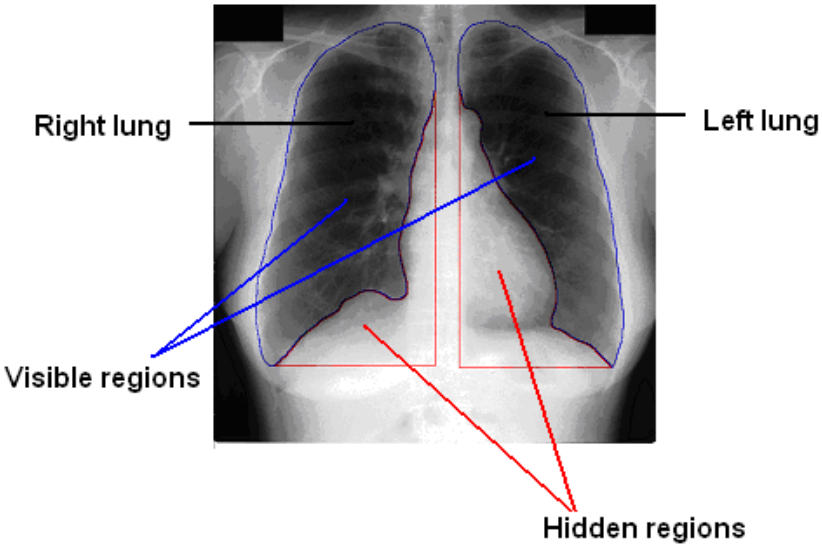


Figure 1.2 - Location of the visible and the hidden regions in a posterior-anterior chest radiograph.

1.2 System Overview

To detect nodules from both visible and hidden regions, we used a CAD system where these regions were assessed separately and the classification results combined in the end.

After segmenting the visible and hidden regions from the chest radiograph, a multi-scale blob detector was used. The objective of the blob detector is to find out bright and roughly circular regions in the radiograph, which are the main characteristics of a nodule. The idea is to locate a set of nodule candidates that will later be classified as a nodule or non-nodule. The next stage, feature measurement, aims to compute a set of suitable features which depict the nodule characteristics in order to provide the classifier with useful and meaningful information. A good set of features is important to reach good results in the final stage: classification. The main objective of classification lies in discarding as many false positive candidates (non-nodule blobs) as possible, without sacrificing too many true positives (nodule blobs). The block diagram of the complete CAD system is shown in figure 1.3.

All the algorithms described in this work were developed in Matlab® (version 7.5.0.342, R2007b)

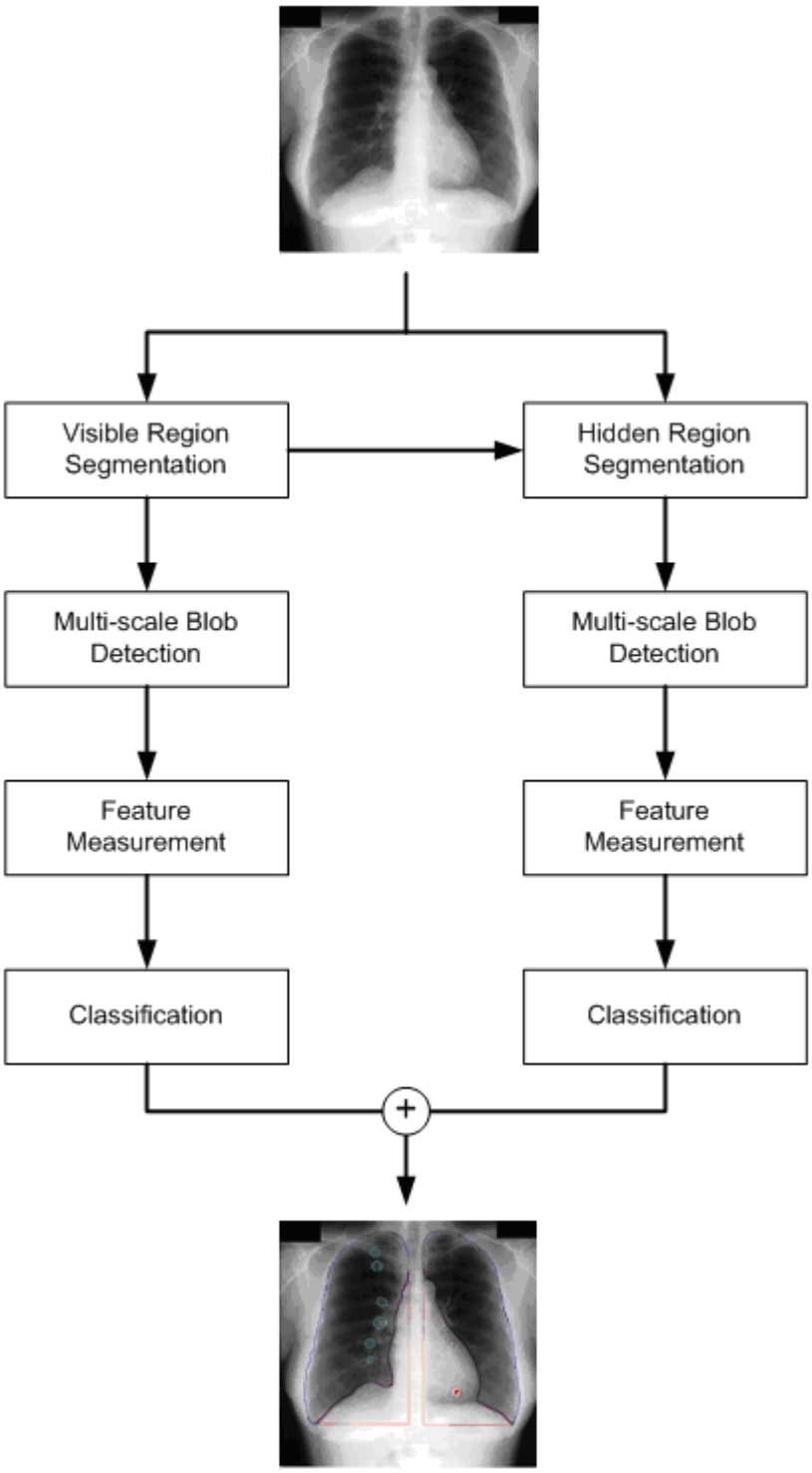


Figure 1.3 - Block diagram of the CAD system.

1.3 Contributions

In the following list are summarised the main contributions of this dissertation:

- It is presented a new methodology for the detection of lung nodules in posterior-anterior chest radiographs. The novelty of this methodology lies in the parallel evaluation of nodule candidates in the visible and hidden regions of the lung.
- The lung area considered in this CAD system includes the regions obscured by the mediastinum, the diaphragm and other anatomic structures, which are usually excluded from the methods presented in the literature.
- A novel combination of a multi-scale blob detector and a Support Vector Machine was used for the visible region.
- An improved version of the multi-scale blob detector is presented, where a sensitivity parameter was introduced to reduce the number of false positives per image in this stage.
- During the classification stage in the visible region, *1-Norm Soft Margin Support Vector Machines* were employed and to deal with imbalanced datasets (the amount of non-nodule candidates was far greater than the amount of nodule candidates), a new strategy based on *box constraints* was used.
- For the hidden region, a new classifier – *Hyper-polyhedron Classifier* (one-class classifier with a very low computational effort) – was conceived and used.

1.4 Document Outline

Chapter 2 (Literature Review) addresses several topics found in the literature: techniques for image enhancement and detection of anatomic structures in chest radiographs, methodologies for selecting the lung nodule candidates and, lastly, most

used classification schemes. Next, in *chapter 3*, the publicly available JRST database is characterised and some preliminary operations are explained, mainly the lung field segmentation of the visible and hidden regions.

The succeeding chapters describe how the next stages of the present CAD system were implemented: multi-scale blob detector (*chapter 4*); feature measurement (*chapter 5*) and classification in the visible and hidden regions (*chapter 6*). Then, experimental results are reported, analysed and discussed in *chapter 7*. Lastly, conclusions are drawn and future work is oriented in *chapter 8*.

At this point we would like to inform that, for visualization purpose, the contrast of almost all images presented in this document was enhanced using the histogram equalization technique.

Chapter 2

Literature Review

In the last decades, many researchers have developed and proposed several methods for detecting lung nodules in posterior-anterior chest radiographs. Typically, after having the lung region identified, the proposed methods can be divided into three main stages: (i) detection of an initial set of nodule candidates; (ii) feature measurement; (iii) classification in order to reduce the number of false positives. Sometimes, it is used a pre-processing step before (i) based on image enhancement or subtraction techniques.

In this chapter, we will organise the literature review by identifying the main contributions in the main phases of the process, namely for image enhancement, nodule candidate detection, feature measurement and classification.

2.1 Gaussian and Laplacian of Gaussian Filters

To begin with, it is important to review some concepts related to Gaussian and Laplacian of Gaussian Filters ([6], [7]) because these filter were used during the development of the multi-scale blob detector of the present CAD system.

Gaussian filters are a class of smoothing filters used in image processing, where the kernel values have a 2-D Gaussian shape. The Gaussian kernel, $G(x, y, \sigma)$, can be given by

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}}$$

where σ is the standard deviation and *(x, y)* are the coordinates of the original image. Notice that the degree of smoothing is parameterised by the standard deviation of the filter. Figure 2.1 illustrates a typical Gaussian kernel (with σ = 8).

Another useful filter in image processing (for example, for edge detection) is the Laplacian of Gaussian (LoG) filter, $H(x, y, \sigma)$. The LoG filter computes, at each point of original image, the sum of the second-order Gaussian derivatives of the original image with respect to *x* and *y* (for a chosen a standard deviation σ), i.e.,

$$H(x, y, \sigma) = \frac{\partial^2 G(x, y, \sigma)}{\partial x^2} + \frac{\partial^2 G(x, y, \sigma)}{\partial y^2}$$

Therefore, we can compute $H(x, y, \sigma)$ as

$$H(x, y, \sigma) = \frac{x^2 + y^2 - 2\sigma^2}{2\pi\sigma^4} \mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (2.1)$$

Figure 2.2 illustrates a typical Gaussian and LoG kernel (with σ = 8). The circular symmetry of the filter often leads to a polar representation, $H(r, \sigma)$, where *r* and *(x, y)* are related by $r = \sqrt{x^2 + y^2}$. In this case, the Laplacian of Gaussian is given by

$$H(r, \sigma) = \frac{1}{\pi\sigma^4} \left(\frac{\mathrm{r}^2}{2\sigma^2} - 1 \right) \mathrm{e}^{-\frac{r^2}{2\sigma^2}}$$

Thus, the zero crossing happens when:

$$H(r, \sigma) = 0 \Rightarrow r = \sqrt{2}\sigma \approx 1.5\sigma \qquad (2.2)$$

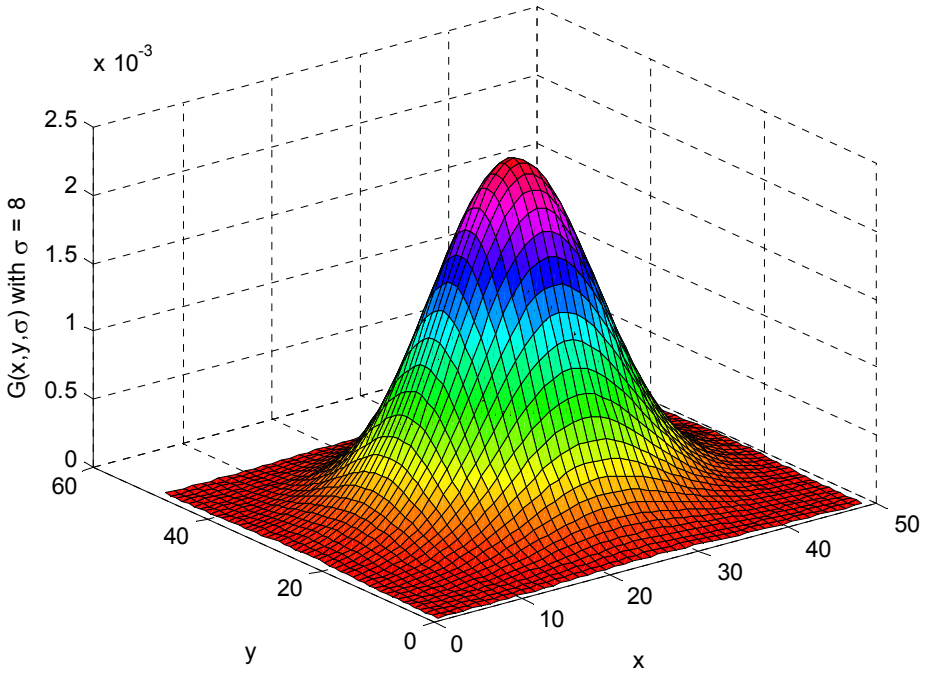


Figure 2.1 – Gaussian kernel, G(x, y, σ), with σ = 8.

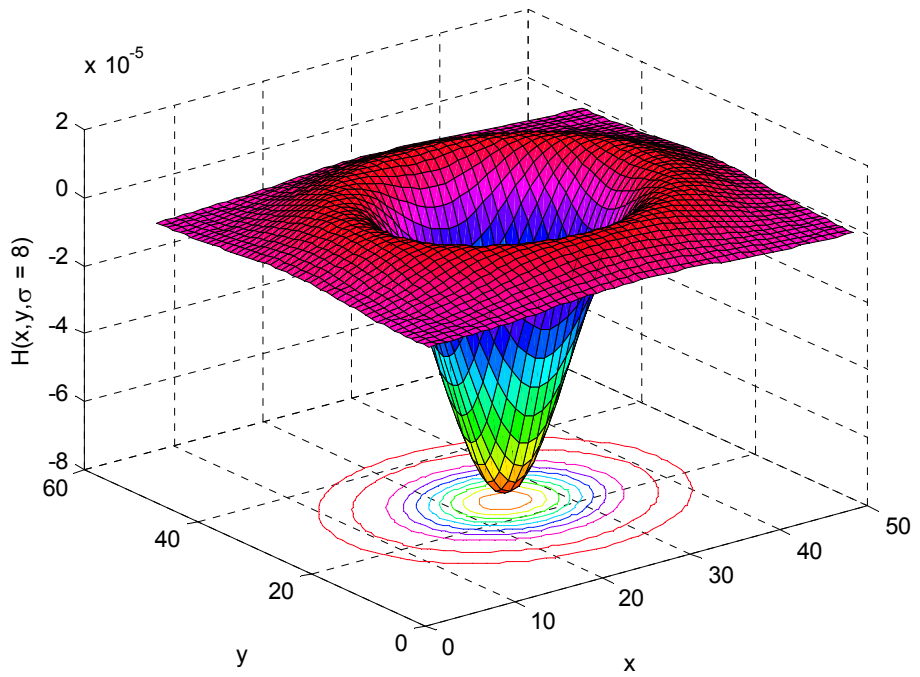


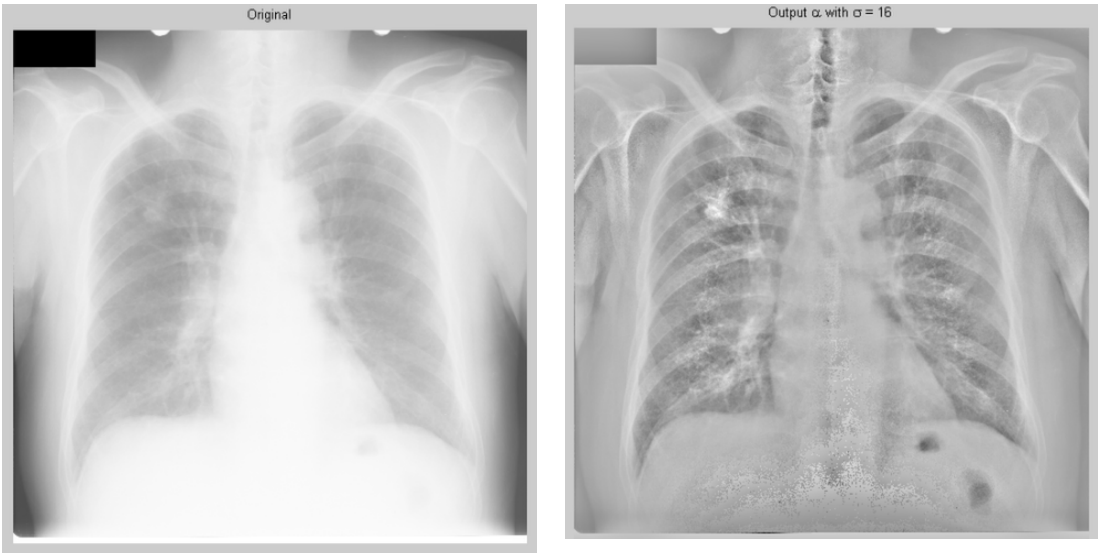
Figure 2.2 – Laplacian of Gaussian kernel, H(x, y, σ), with σ = 8 (the so-called Mexican Hat).

2.2 Image Enhancement

Usually, contrast in chest x-ray images is low and it is often hard to detect a specific kind of abnormality. Thus, some methods use a pre-processing step to increase contrast. In [8] is proposed an enhancement solution based on local normalisation (LN). The output image of this filter, $\boldsymbol{I_{LN}}$, is obtained using the following expression:

$$I_{LN} = \frac{I - \bar{I}}{\sqrt{(\overline{I^2}) - (\bar{I})^2}}$$

where, $\bar{I}$ is the result of the convolution between the original image, $\boldsymbol{I}$, and a Gaussian kernel. Figure 2.3 (b) shows the result obtained after applying the local normalization filter (σ = 16) to a chest radiograph. The result is a more contrasted image with a noticeable image enhancement.



(a) (b)

Figure 2.3 - (a) Original chest radiograph; (b) The result of applying the local normalization filter (σ = 16).

Other methods explore the idea of some lung regions are obscured by anatomic structures. Thus, subtraction techniques attempt to remove normal structures in chest

radiographs so that abnormalities stand out more clearly.

An approach is temporal subtraction whereby an input image is compared with a previous radiograph of the same patient. For instance, Zhao et al. [9] proposed a temporal subtraction technique that performs elastic registration[†] on rib border segments detected with an adaptive oriented filter. The authors reported that the contrast around lung nodules substantially increased in the subtracted images.

If a previous radiograph is not available, Li et al. [10], [11] proposed a subtraction technique which makes use of the symmetry of the rib cage. This subtraction can be made by mirroring the left/right lung field, performing elastic registration on the right/left lung field and subtracting.

Another example is the automatic detection of rib borders proposed by Moreira et al. [12] whereby a geometric model described as a second-degree polynomial is fitted along the ribs.

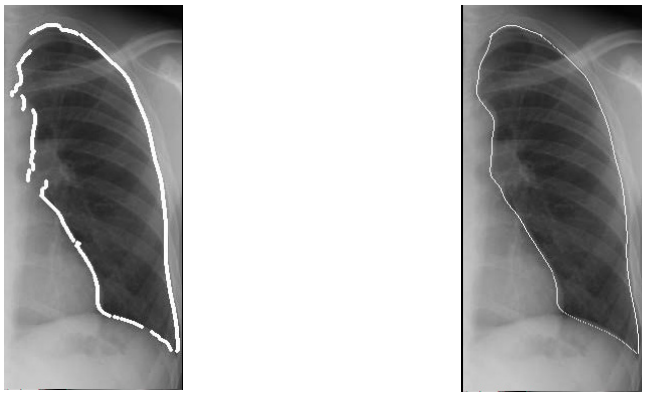
2.3 Lung Field Segmentation

The lung field delimitation is a necessary step before starting to look for nodules. It is not a simple problem and several systems using distinctive methodologies have been proposed. Some of them are presented and compared in [13]. Recently, the active shape model (ASM) methodology has been used in several CAD systems related to lung nodule detection (for instance, [8] and [14]). In the present CAD system the approach proposed by Mendonça et al. [15] will be employed (see also [16]). The sequence of main operations used in this methodology can be summarised in the following steps:

1. Detection of mediastinal and costal edge pixels using 1-D first derivative Gaussian filters to image rows.

[†] Nonlinear alignment process by which two images (or image regions) of the same set of objects are positioned coincident with each other. It is used to match two or more images taken at different times.

2. Detection of top and bottom edge pixels using 1-D first derivative Gaussian filters to image columns.
3. Combination of detected edge pixels to define a closed contour using a spline approximation method. Figure 2.4 (b) shows the final contour for a left lung.



(a) (b)

Figure 2.4 - (a) edge pixels detected along the lung border; (b) border approximation by splines (these images were taken from [15]).

2.4 Nodule Candidate Detection

The objective of the nodule candidate detector is to find out bright and/or roughly circular regions in the original image (radiograph), which are the main characteristics of a nodule. The idea is to locate a set of nodule candidates that will later be classified as a nodule or non-nodule.

Some methods employ the difference-image technique to candidate detection ([17], [18]). The main idea is firstly to filter the original image with a spherical kernel to obtain a nodule-enhanced image. Secondly, the original image is filtered with a median filter to obtain a nodule suppressed image. Nodule candidates are then obtained by thresholding the subtraction of the two filtered images.

Other methods employ Convergence Index (CI) filters or similar filters ([19], [20], [21]). These filters are used to detect rounded convex regions. For each pixel of interest,

P, the evaluation is done in a pre-defined region of support around *P*, where the filter evaluates the degree of convergence of gradient vectors[‡]. The basic idea is that structures in the image with circular shape that are brighter in the inner region will tend to have gradients that point towards the centre of the object. This would yield a high CI value. Remarkably, the CI value depends on the gradient direction and not the gradient magnitude. This can be useful for detecting very dim and subtle nodules.

A different technique is proposed by Ginneken et al. [8] which is based on scale-space representation introduced by Lindeberg [22]. The idea is to apply (normalised) Laplacian of Gaussian (LoG) filters at several scales, σ, in order to stand out bright and roughly circular regions with different sizes. A similar technique was used in this work and a detailed description of this method is presented in chapter 4.

2.5 Feature Measurement

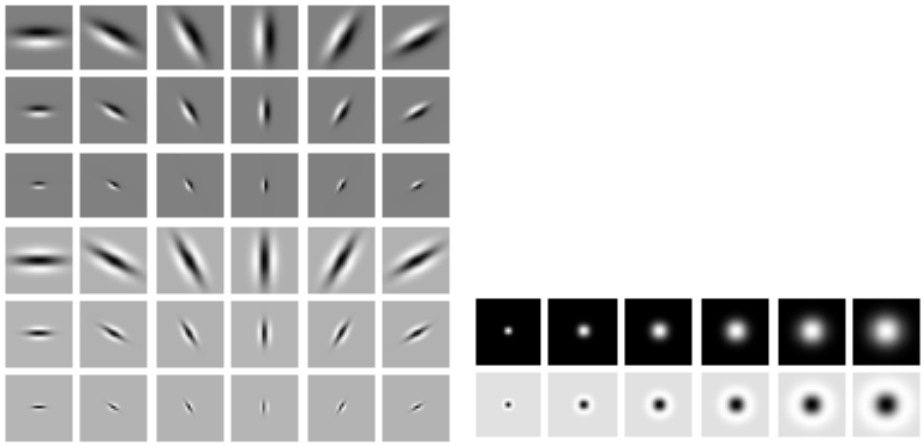
The objective of feature measurement is to provide the classifier with a set of measures from each nodule candidate, and consequently, a good set of features can lead to good results in the classification stage.

There is a wide range of meaningful characteristics for candidate classification used in published methods. A list and description of some of these features can be found in [5], [14], [23] and [24]. These features can be sorted into three main groups:

- *Geometric features* (e.g. estimated radius / size of suspicious region, circularity, position and distance to lung perimeter);
- *Intensity and Contrast features* (e.g. mean intensity of suspicious region; centre-boundary contrast);
- *Features derived from first and second-order filters at different scales and directions* (e.g. gradient magnitude, radial gradient, LoG filters).

[‡] The gradient of the image intensity is the vector: $\nabla \mathbf{f} = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right]^t$, where *f(x, y)* is the intensity at each point *(x, y)* of the original image.

Some authors have also proposed the use of texture analysis methodologies to extract characteristics. For instance, Pereira et al. [25] propose a scheme based on a MR8 (Maximum Response) filter bank, which was extended to six Gaussian filters and six LoG filters at different scales. The MR8 filter bank also comprises edge filters at 3 scales and bar filters at the same 3 scales. The last two sets also occur at 6 orientations at each scale. The proposed filter bank is used to extract features from suspicious region. The kernels of these filters are presented in figure 2.5.



(a) (b)

Figure 2.5 - Kernels of the proposed filter bank in [25]: (a) The MR8 filter bank; (b) Gaussian and LoG filters at different scales.

2.6 Classification

The last step is the candidate classification which has the objective of discarding as many false positive candidates as possible, without sacrificing too many true positives. Some methods have been proposed, such as learning systems based on neural network (NN) classifiers ([18], [26], [27]), K-nearest-neighbour (KNN) classifiers ([8]) or Support Vector Machines ([28], [29]).

Ginneken et al. [8] proposed a scheme based on two consecutive classification steps. Both steps make use of K-nearest-neighbour (KNN) classifiers [30]. The first step aims

to reduce the large number of candidates using a reduced set of features (13 features). Then, in the second step, it is extracted the final subset using a large set of features (109 features).

Campadelli et al. [28], [29] present a different methodology in which Support Vector Machines (SVMs) are employed for candidate classification. They use several polynomial and Gaussian kernels. In fact, in the field of medical image processing, several authors (for instance, [31] - [35]) have presented CAD systems which employ SVMs for candidate classification.

2.6.1 Support Vector Machine Brief Overview

As SVMs were employed in the present CAD system for the visible region, we introduce here this methodology. A detailed description of SVMs can be found in [36], [37], [38].

Considering a set of *l* training samples, $S = \{(\mathbf{x}_1, y_1),...,(\mathbf{x}_l, y_l)\}$. We refer to the $\mathbf{x}_i$ as samples (feature vectors) and the y_i as class labels, where $y_i = \{-1,1\}$ for binary classification.

Let us analyse the simplest case: a hard margin linear SVM. It should be noticed that the hard margin classifier works only for data which are linearly separable in the feature space. The linear decision function is given by $f(\mathbf{x}) = \mathbf{w}' \cdot \mathbf{x} + b$ ($\mathbf{w}'$ denotes the transpose of $\mathbf{w}$) and, geometrically, the feature space is split into two parts by the separating hyper-plane defined by the equation $f(\mathbf{x}) = 0$ (figure 2.6 (a)).

The aim of the SVM is to work out the separating hyper-plane that maximises the value of the geometric margin, γ, which is equal to half the distance between the dashed lines in figure 2.6 (a). Notice that the points that lie closest to the hyper-plane, i.e., the points that lie along the dashed lines in figure 2.6 (a) are the *support vectors*.

Considering that $\mathbf{w}' \cdot \mathbf{x} + b = -1$ and $\mathbf{w}' \cdot \mathbf{x} + b = 1$ for the support vectors from class C_1 ($\mathbf{x}_{svC1}$) and class C_2 ($\mathbf{x}_{svC2}$) respectively (figure 2.6 (b)), the geometric margin is

given by

$$\gamma = \frac{1}{2}\frac{(\mathbf{w}' \cdot \mathbf{x}_{svC1} + b) - (\mathbf{w}' \cdot \mathbf{x}_{svC2} + b)}{\|\mathbf{w}\|_2} = \frac{1}{\|\mathbf{w}\|_2}$$

Thus, the maximisation of the geometric margin is equivalent to the minimisation of $\|\mathbf{w}\|_2$. Therefore, solving the following optimisation problem, we obtain the maximal margin hyper-plane.

$$\min_{\mathbf{w},b} = \mathbf{w}' \cdot \mathbf{w}$$

subject to:

$$y_i(\mathbf{w}' \cdot \mathbf{x}_i + b) \geq 1$$
$$i = 1,...,l$$

To compute the solution, we can convert the above optimisation problem into the following quadratic optimisation problem [36]:

$$\max w(\boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i' \cdot \mathbf{x}_j)$$

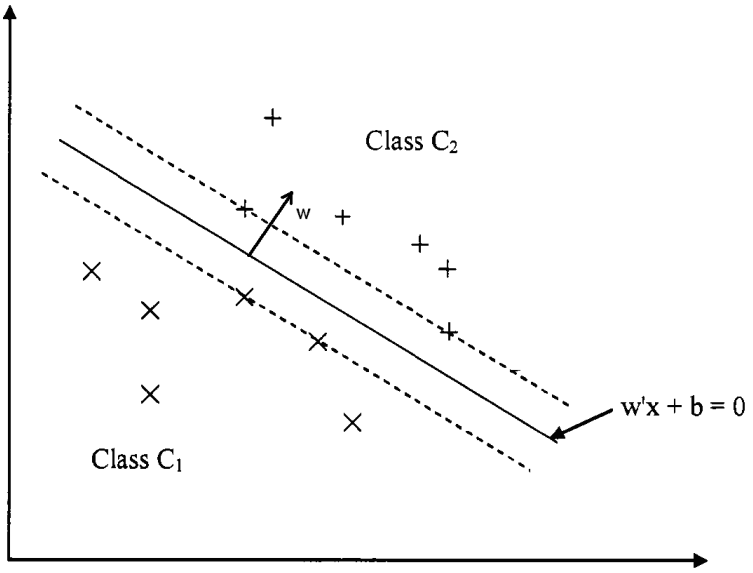
subject to:

$$\sum_{i=1}^{l} \alpha_i y_i = 0$$
$$\alpha_i \geq 0, i = 1,...,l$$

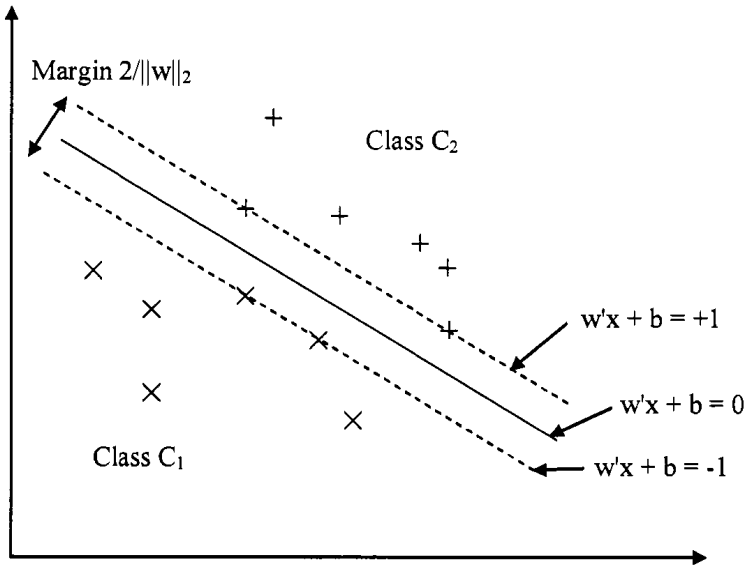
Then, the decision rule is given by the sign of $f(\mathbf{x})$, where

$$\text{sign}[f(x)] = \text{sign}\left[\sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i' \cdot \mathbf{x}) + b\right]$$

The subset of points that lie closest to the hyper-plane (their corresponding α_i is non-zero) contains all the information necessary to reconstruct the hyper-plane. It is for this reason that these points are called *support vectors*. Even if all of the other points (non-support vectors) were removed, the same separating hyper-plane would be found for the remaining subset of the support vectors.



(a)



(b)

Figure 2.6 – (a) separating hyper-plane for a two dimensional training set; (b) the distance between the dashed lines (which contain the support vectors) is twice the geometric margin, i.e., $2\gamma = 2/\|\mathbf{w}\|_2$.

Chapter 3

The JRST Database and Preliminary Operations

In this chapter, we begin by describing the JRST (Japanese Society of Radiological Technology) database. Next, we present the evaluation method used for assessing the present CAD system. Finally, we describe the lung field segmentation methodology used in this system.

3.1 The JRST Database

The present CAD system was evaluated in chest radiographs from the publicly available JSRT (Japanese Society of Radiological Technology) database, described in [4]. We used this database because of its frequent use in other research work (for instance, [8], [14], [29]), which allows us the comparison of the performance.

The database contains 247 posterior–anterior chest radiographs of which 154 are patients with lung nodule (a sample of them is shown in figure 3.1) and 93 are normal cases without nodules. All the images were assessed by radiologists and the diameter

and position of the nodules is provided. These nodules were graded according to subtlety, pathology (malignant or benign), and size. The subtlety rating includes five levels: extremely subtle, very subtle, subtle, relatively obvious, and obvious, as summarised in table 3.1.

It should be notice that 141 out of 154 nodules are located in the visible region, whereas 13 nodules are located in the hidden region (they represent about 8.5% of the total number of nodules in the JRST database).

Table 3.1 - Distribution of nodules in the JRST database stratified by subtlety and size (this information was taken from [14]).

Subtlety	**Size**			**Total**
	Small (<10 mm)	**Medium (≥ 10 mm and <20 mm)**	**Large (≥ 20 mm)**	
Extremely subtle	2	18	5	25 (16.2%)
Very subtle	3	16	10	29 (18.8%)
Subtle	4	29	17	50 (32.5%)
Relatively obvious	1	20	17	38 (24.7%)
Obvious	0	5	7	12 (7.8%)
	(6.5%)	(57.1%)	(36.4%)	

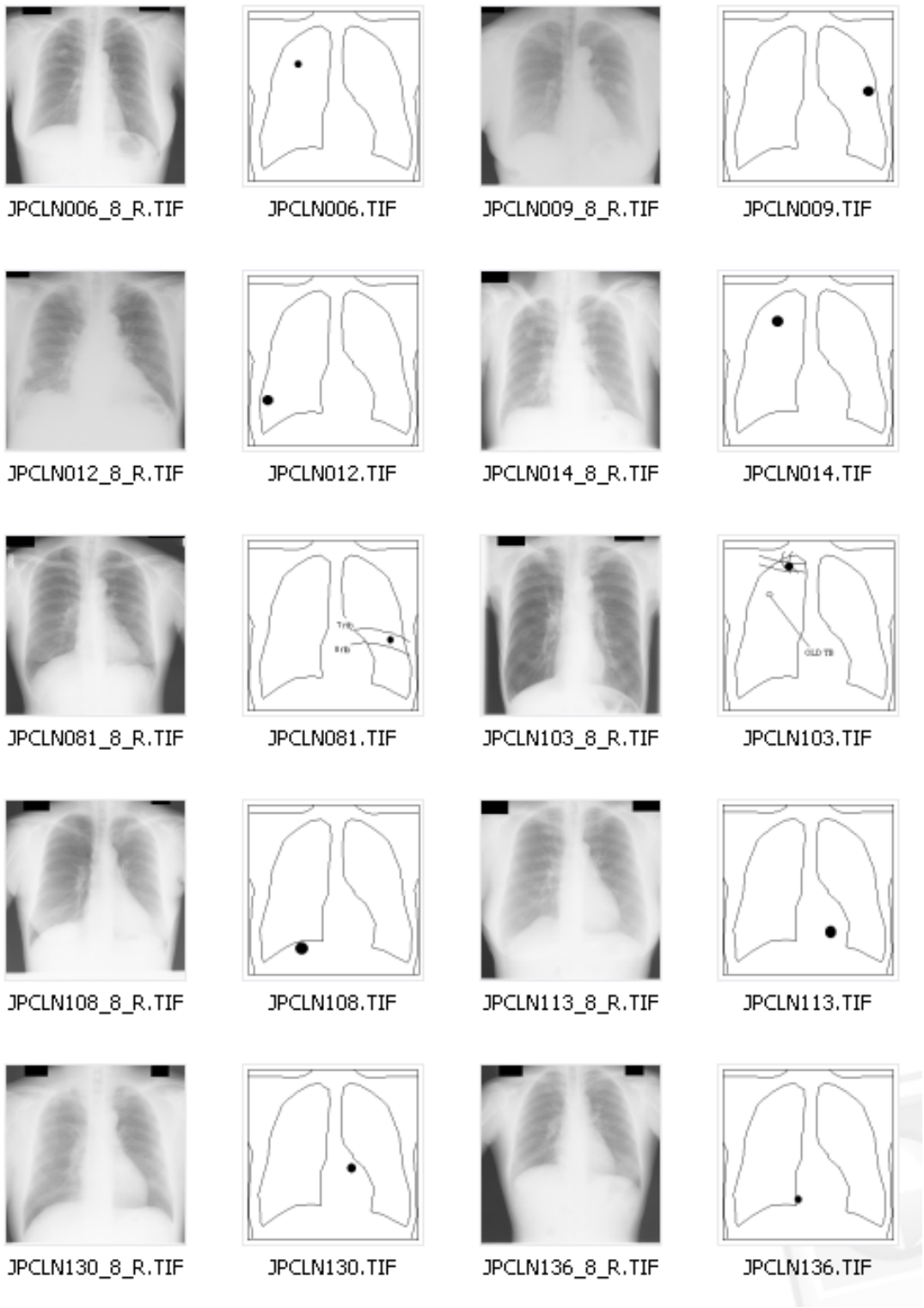


Figure 3.1 - Several chest radiographs from the JRST database and respective positions of nodules.

The JSRT database images have been digitised to 2048 × 2048 pixels and the size of one pixel is 0.175 × 0.175 mm^2. Before using these images, they were down-sampled to 512 by 512 pixels which made the size of each pixel be 0.7 × 0.7 mm^2 and reduced the computational effort needed to process each image.

3.2 Evaluation Method for CAD System

The present CAD system was evaluated in terms of sensitivity and the number of false positives per image (FP per image). Notice that in this document "*per image*" means "*per patient*" (including right and left lungs).

The sensitivity [39] measures the number of correctly classified candidates out of the number of all nodules and is given by,

$$sensitivity = \frac{TP}{TP + FN}$$

Another performance measure is specificity [39], which is given by,

$$specificity = \frac{TN}{TN + FP}$$

In this context,

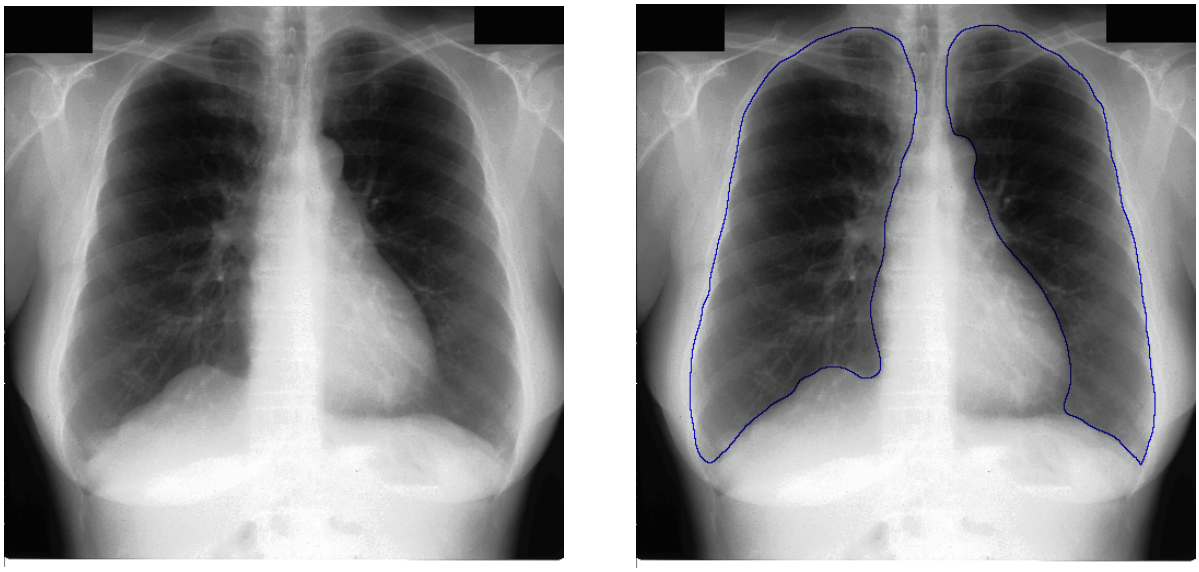
- *TP* denotes the number of true positives, i.e., suspicious regions correctly labelled as nodules;
- *FP* denotes the number of false positives, i.e., suspicious regions wrongly labelled as nodules;
- *TN* denotes the number of true negatives, i.e., suspicious regions correctly labelled as non-nodules.
- *FN* denotes the number of false negatives, i.e., suspicious regions wrongly labelled as non-nodules;

3.3 Lung Field Segmentation

For automated segmentation of the lung visible region, we can employ the approach proposed by Mendonça et al. [15] (see also [16]). As mentioned in chapter 2, the sequence of main operations used in this methodology can be summarized in the following steps:

1. Detection of mediastinal and costal edge pixels using 1-D first derivative Gaussian filters to image rows.
2. Detection of top and bottom edge pixels using 1-D first derivative Gaussian filters to image columns.
3. Combination of detected edge pixels to define a closed contour using a spline approximation method.

In figure 3.2 is presented a typical case with the lung visible regions outlined.



(a) (b)

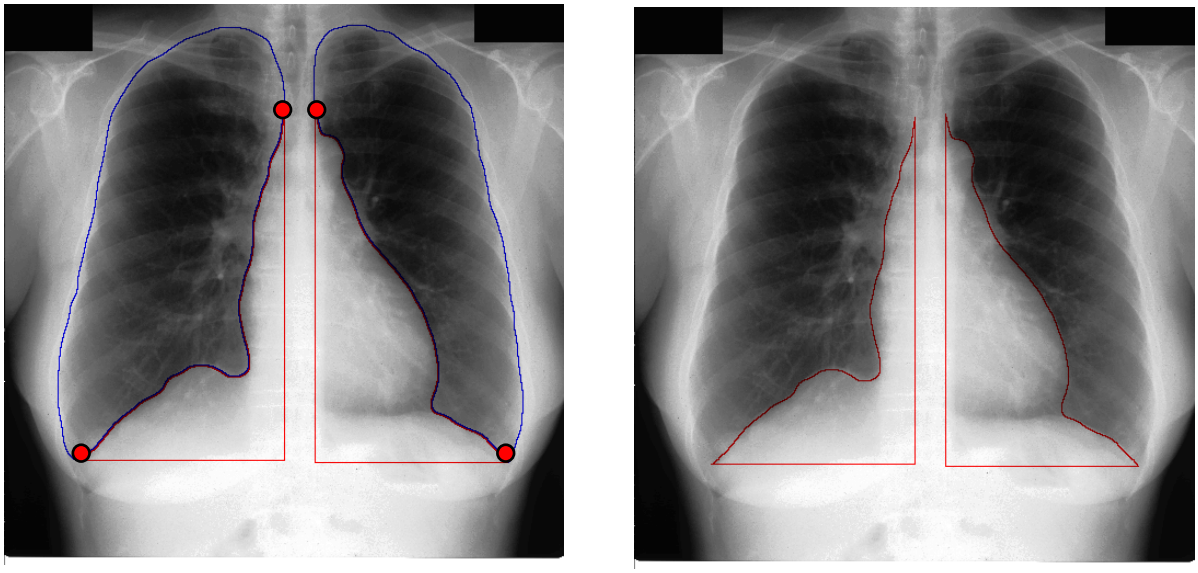
Figure 3.2 - Chest radiograph (case JPCLN113) with the visible lung region outlined (b).

However, a significant part of the lungs is obscured by the mediastinum, the diaphragm and other anatomic structures. Consequently, if a nodule was located in those parts of the lung, it would be missed by detection systems that consider only the lung visible region. Hence, to extend the region of interest in order to include those hidden regions, a simple algorithm that automatically delineates them was developed.

Firstly, the algorithm finds out four key points:

- In the right lung visible region: the bottom right corner and the far left point.
- In the left lung visible region: the bottom left corner and the far right point.

Then, using these points, two perpendicular lines per lung (one vertical and one horizontal) are used to delineate the boundaries of the hidden region. These boundaries are illustrated in figure 3.3.



(a) (b)

Figure 3.3 – (a) chest radiographs (case JPCLN113) with the four key points highlighted; (b) chest radiographs with only the hidden lung region outlined.

Chapter 4

Nodule Candidate Detection

The objective of the nodule candidate detector is to find out bright and roughly circular regions in the original image (radiograph), which are the main characteristics of a nodule. The idea is to locate a set of nodule candidates that will later be classified as a nodule or non-nodule.

4.1 Multi-scale Blob Detector

In chest radiographs, nodules appear typically as bright and roughly circular blobs. As discussed previously, there is a wide range of nodule sizes. For instance, the nodules in the JSRT database range in diameter from 6 to 60 mm [8]. Consequently, a suitable approach to deal with this wide range of diameters can be multi-scale detection. As the multi-scale blob detector methodology used by Ginneken et al. [8] (based on scale-space representation) revealed promising results, a similar methodology was employed here. However, to improve this method, especially to reduce the number of FPs per image, we introduced some changes described below, such as the introduction of a

sensitivity parameter.

The basic idea of scale-space representation is to describe the original image data at multiple scales ([22]). Therefore, it is performed the convolution (denoted as $*$) between the original image, $f(x,y)$, and the Gaussian kernels, $G(x,y,\sigma)$, with successively greater standard deviation, σ.

$$L(x,y,\sigma) = G(x,y,\sigma) * f(x,y)$$

where $L(x,y,\sigma)$ denotes the scale-space representation at scale σ. The σ is referred as the scale parameter.

Firstly, <u>for each scale σ</u>, we applied to the original image (radiograph), *f(x,y)*, a Laplacian of Gaussian (LoG) kernel with standard deviation σ, $H(x,y,\sigma)$. The idea is to calculate, at each point of the original image, the sum of the second-order Gaussian derivatives of $f(x,y)$ at scale σ, with respect to x and y, i.e., the Laplacian of $L(x,y,\sigma)$, $\nabla^2 L(x,y,\sigma)$. We added a minus sign so that in an image intensity transition (for instance, in the edge of a nodule) we had a positive maximum in the brighter region and not a negative minimum. Thus, we defined $\nabla^2 L(x,y,\sigma)$ as

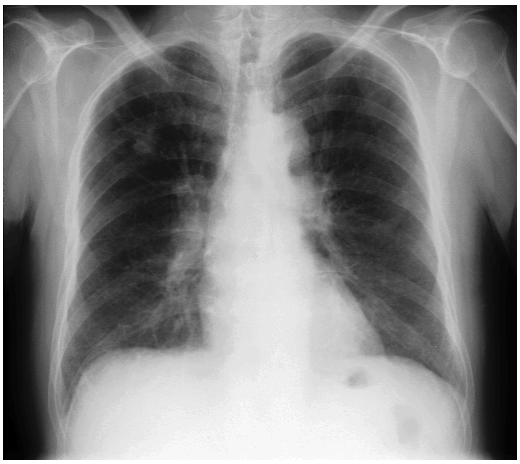
$$\nabla^2 L(x,y,\sigma) = -\left(\frac{\partial^2 L(x,y,\sigma)}{\partial x^2} + \frac{\partial^2 L(x,y,\sigma)}{\partial y^2}\right) = -H(x,y,\sigma) * f(x,y)$$

However, the amplitude of scale-space representation decreases exponentially with scale (see equation 2.1, chapter 2) and consequently, it was necessary to use the factor σ^2 which allowed for comparison of the Laplacian output at different values of σ. Thus, the normalised Laplacian, $\nabla^2{}_{norm} L(x,y,\sigma)$, is given by

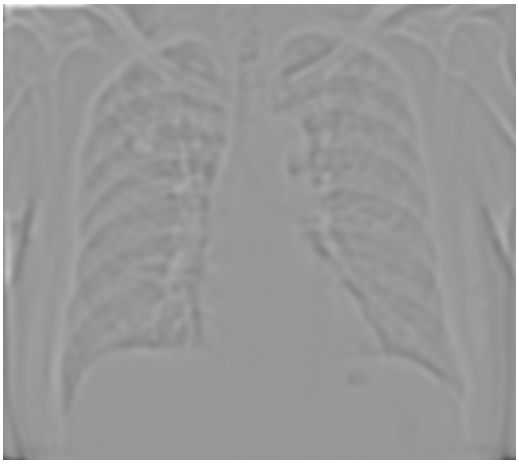
$$\nabla^2{}_{norm} L(x,y,\sigma) = \sigma^2 \times \nabla^2 L(x,y,\sigma)$$

The original image blob candidates were located in the local maxima of $\nabla^2{}_{norm} L(x,y,\sigma)$ for each scale σ.

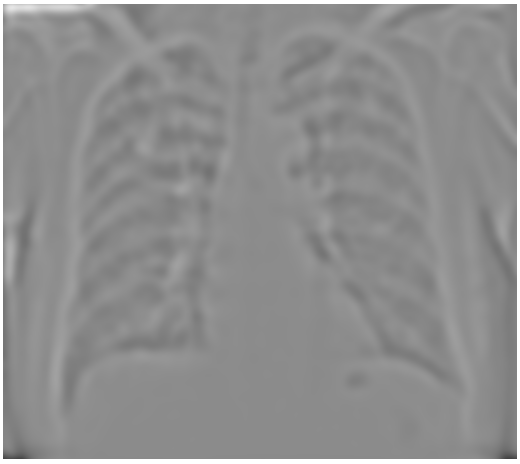
Figure 4.1 shows the results of applying the normalised Laplacian at different scales.



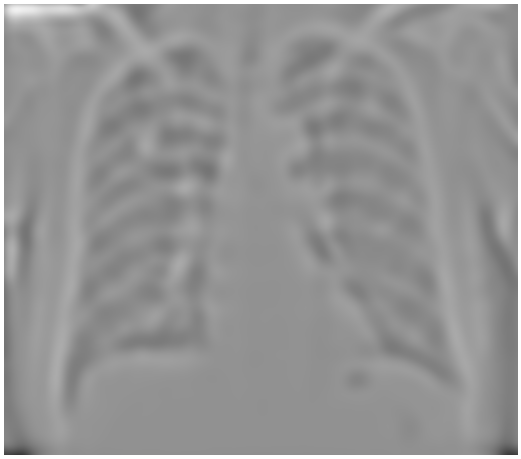
(a)



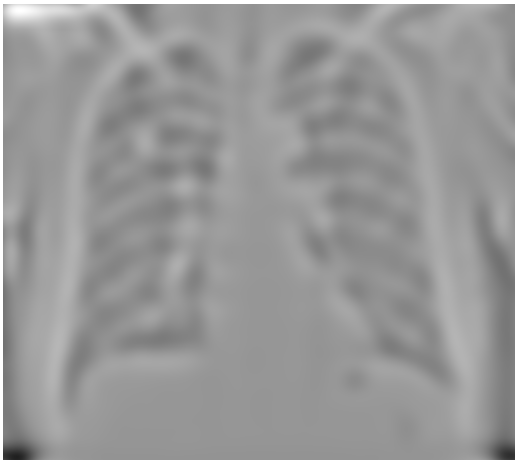
(b)



(c)



(d)



(e)

Figure 4.1 - Original image (a) and the results of applying $\nabla^2_{norm} L(x, y, \sigma)$ for: σ = 4 (b); σ = 6 (c); σ = 8 (d); σ = 10 (e).

To reduce the number of blobs, we introduced a sensitivity parameter ξ, and only blobs stronger than ξ were considered, i.e., blobs that satisfy the equation:

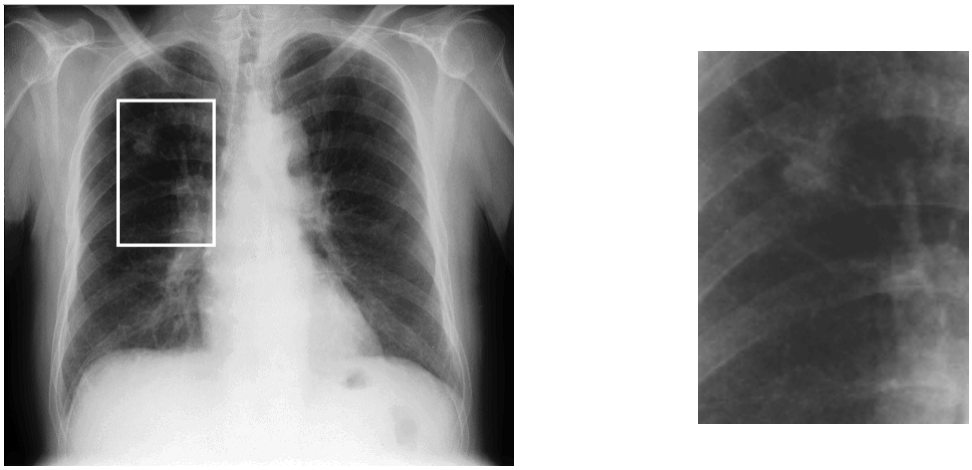
$$\nabla^2{}_{norm} L(x_C, y_C, \sigma) \geq \xi$$

where x_C and y_C denote the blob centre coordinates.

In the next stage, the normalised Laplacian, $\nabla^2{}_{norm} L(x, y, \sigma)$, was used for blob strength comparison. When blobs from different scales (or at the same scale) were overlapped only the blob with the highest value of $\nabla^2{}_{norm} L(x_C, y_C, \sigma)$ was kept.

To illustrate these steps, let us zoom in on the original radiograph as shown in figure 4.2. Figure 4.3 shows all blobs detected only at a specific scale and the final result after selecting the strongest blobs.

Finally, for each remaining blob, the position in the image (blob centre coordinates), its corresponding scale of detection σ and the value of $\nabla^2{}_{norm} L(x_C, y_C, \sigma)$ were stored. The approximate blob radius, *r*, can be obtained by multiplying the corresponding scale by the factor 1.5, i.e., $r \approx 1.5\sigma$ (see equation 2.2, chapter 2)



(a) (b)

Figure 4.2 - The original radiograph (a) was zoomed in on the specified region (b).

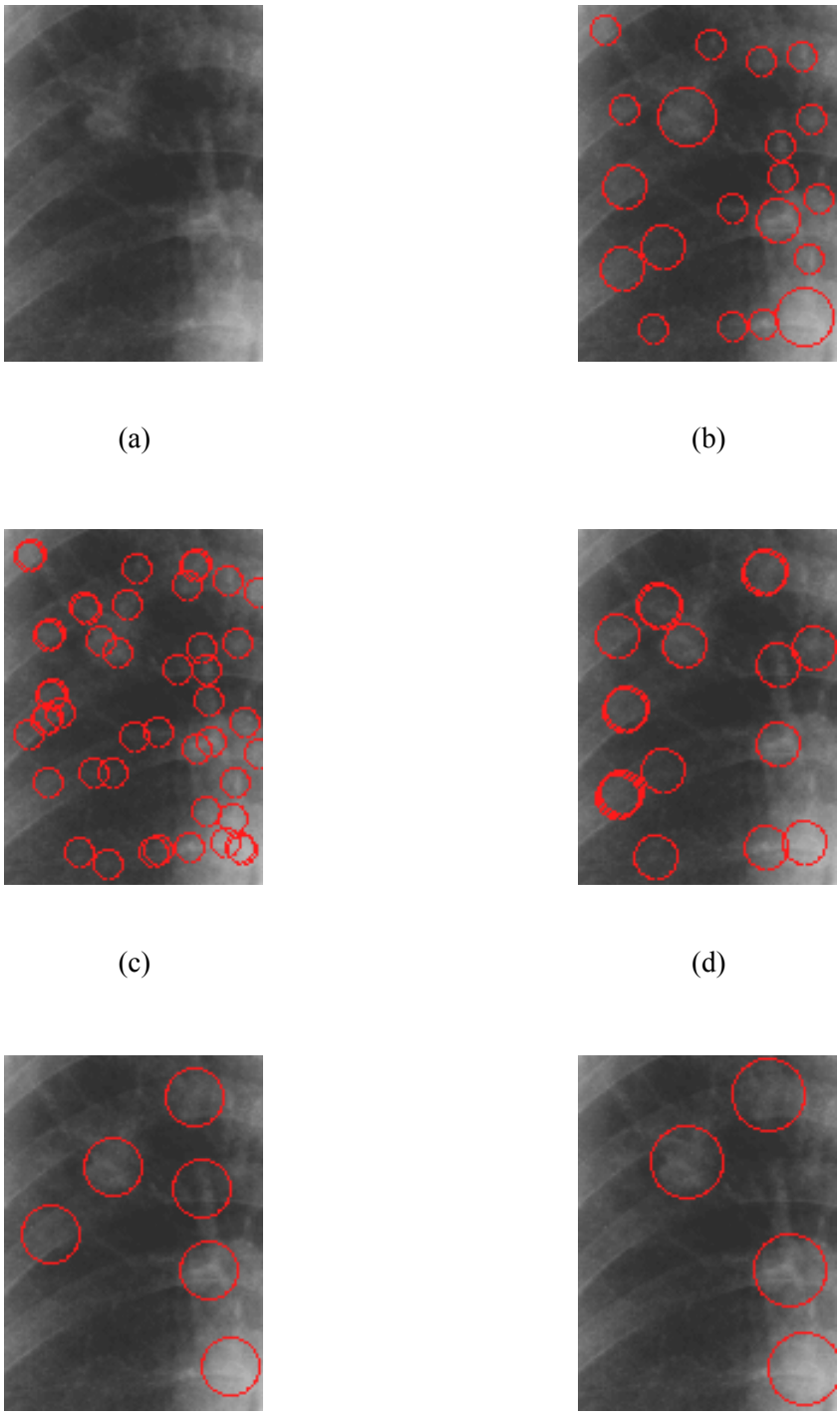
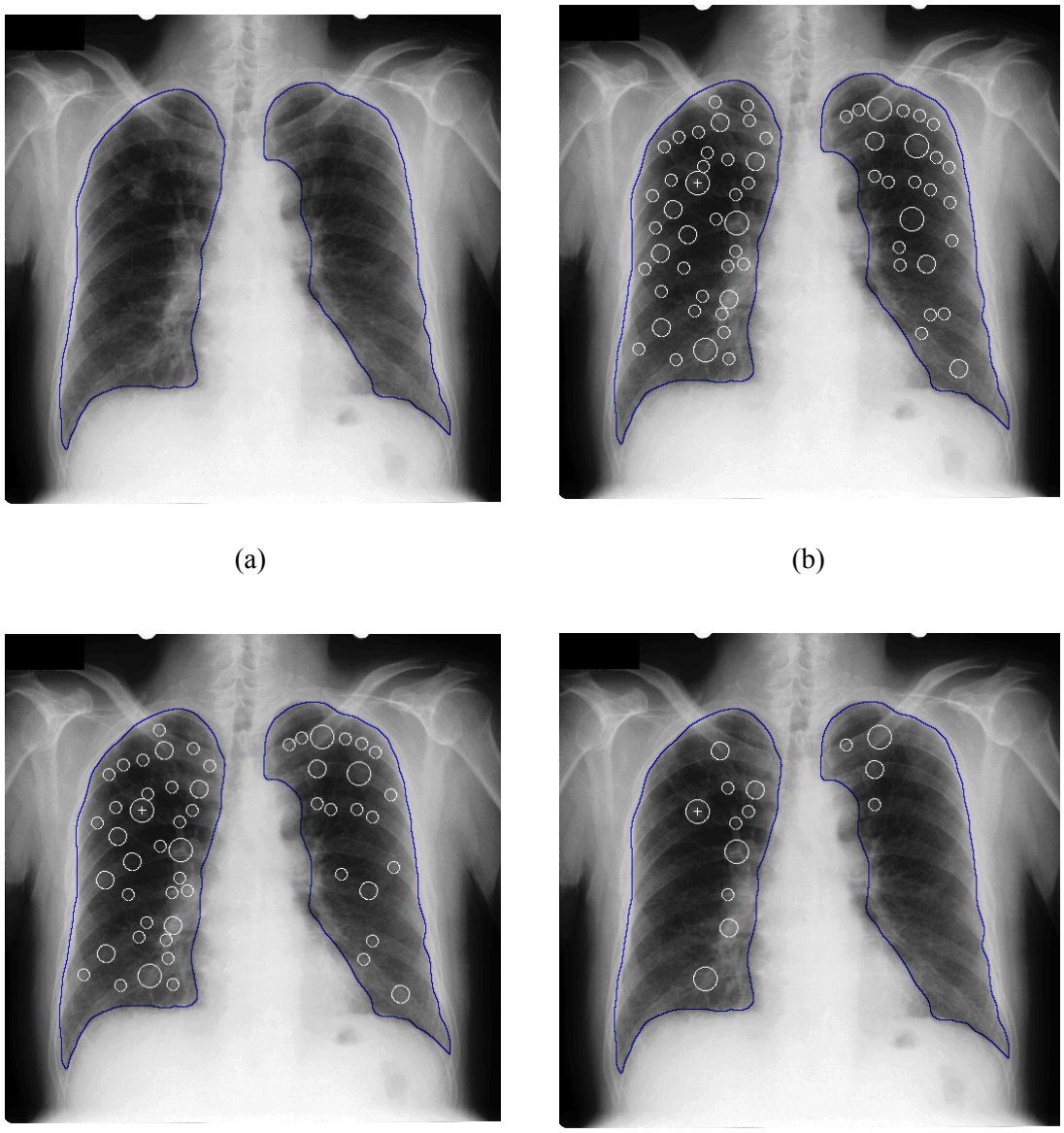


Figure 4.3 - (a) input image; (b) output image after selecting the strongest blobs and all blobs detected only at scale: $\sigma = 4$ (c); $\sigma = 6$ (d); $\sigma = 8$ (e); $\sigma = 10$ (f) ($\xi = 0.01$).

Figure 4.4 illustrates the decrease in the number of blobs detected when we increase the value of parameter ξ.



(a) (b)

(c) (d)

Figure 4.4 - Detected blobs for different values of ξ. (a) the original image; (b) ξ = 0.008 with 63 blobs detected; (c) ξ = 0.01 with 53 blobs detected; (d) ξ = 0.02 with 14 blobs detected (σ={4, 6, 8, 10}).

4.2 Labelling Method

It is important to decide how a nodule candidate should be labelled as a true positive (TP). In [8] (Guiniken et al), a candidate is labelled as a TP if there is an overlap (not necessarily a total overlap) between the truth circle and the blob circle given by the blob detector. However, we were more restrictive and considered a candidate as a TP only if the blob circle includes the real centre of the nodule.

Notice that the problem of detecting a small nodule at a big scale rarely happens. If the nodule was detected at that scale, it was also detected at smaller scales with a greater value of $\nabla^2_{norm} L(x_C, y_C, \sigma)$ and consequently, only the most suitable scale is kept.

4.3 Concluding Remarks

The next step before using the blob detector is the choice of scales and the value of the parameter ξ that lead to good results at the blob detector output. This analysis is done in chapter 7.

Moreover, having detected the bobs in the radiograph, it is necessary to classify them as nodule or non-nodule because several anatomic structures in the lung field have a like blob appearance. Therefore, feature measurement and classification are the next stages.

Chapter 5

Nodule Candidate Features

This stage aims to compute a set of suitable features which depict the nodule characteristics in order to provide the classifier with useful and meaningful information. A good set of features is important to reach good results in the classification stage.

5.1 Intensity and Contrast Features

As the regions near the centre of the nodule tend to be brighter than others close to the boundary, it is useful to measure the *contrast between the central region and the boundary (CBC)* which is given by the difference between the *mean intensity in the central region (CI)* and the *mean intensity along the boundary (BI).* The central region is always made up of the nine central pixels and no more in order to avoid an overlap between the central region and the boundary for blobs at smallest scales. The boundary of the blob is made up of the pixels crossed by the circle with radius $r \approx 1.5 * \sigma$ (see chapter 4).

On the other hand, in the inner region (region inside the boundary of the blob)

several measures to characterise the intensity distribution were computed: *maximum value of intensity inside (maxI)*, *mean intensity inside (μ_I)* and the *standard deviation of the intensity distribution inside (σ_I)* which is important to measure the dispersion of the intensity (I) distribution in the inner region (N pixels) and is given by,

$$\sigma_I = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(I_j - \mu_I\right)^2}$$

It was also computed the *intensity skewness inside (s_I)* and the *intensity kurtosis inside (k_I)*. The *intensity skewness inside (s_I)* measures how asymmetrical is the intensity distribution in the inner region and is given by,

$$s_I = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{I_j - \mu_I}{\sigma_I}\right)^3$$

Finally, the *intensity kurtosis inside (k_I)* measures the peakedness or flatness of the intensity distribution in the inner region and is given by,

$$k_I = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{I_j - \mu_I}{\sigma_I}\right)^4$$

Thus, the more peaked (less flat) is the intensity distribution, the greater is the kurtosis value.

Notice that the intensity skewness inside and the intensity kurtosis inside are related to third (m_3) and fourth (m_4) moments of the intensity distribution, i.e.,

$$s_I = \frac{m_3}{\sigma^3} \quad \text{and} \quad k_I = \frac{m_4}{\sigma^4}$$

5.2 Features derived from first and second-order filters

As nodules are roughly circular and are brighter in the inner region, they will tend to

have gradients that point towards the centre of them. Consequently, it is useful to measure features related to the convergence coefficient and the radial gradient of the nodule candidates.

Gradient:

In order to work out the gradient vector, $\boldsymbol{G}(x, y)$, of the image intensity at each point (x, y) of the suspicious region (inner region), the gradient along X direction and Y direction were computed [7]. Let us denote them by $\boldsymbol{G_X}(x, y)$ and $\boldsymbol{G_Y}(x, y)$, respectively. The X and Y axes are drawn in figure 5.1. The gradient magnitude is given by

$$|G(x,y)| = \sqrt{G_X{}^2(x,y) + G_Y{}^2(x,y)}$$

and the gradient angle (phase) with respect to the Y axis is given by

$$\theta_g(x,y) = \arctan\left(\frac{G_X(x,y)}{G_Y(x,y)}\right)$$

For candidate characterisation two measures were computed:

- *Gradient magnitude mean:* $\mu_{|G|} = \frac{1}{N}\sum_{j=1}^{N}|G|_j$

- *Gradient magnitude standard deviation:* $\sigma_{|G|} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(|G|_j - \mu_{|G|}\right)^2}$

Convergence coefficient:

The convergence coefficient is a measure of how strongly the gradient vector points to the blob central pixel. For instance, if the gradient vector points exactly to the centre, then the convergence coefficient is 1 (maximum value).

The convergence coefficient at pixel P (see figure 5.1) can be defined as $cos(\theta_r)$ where θ_r is the angle between the radial vector pointing from pixel P to the centre C and the gradient vector at P.

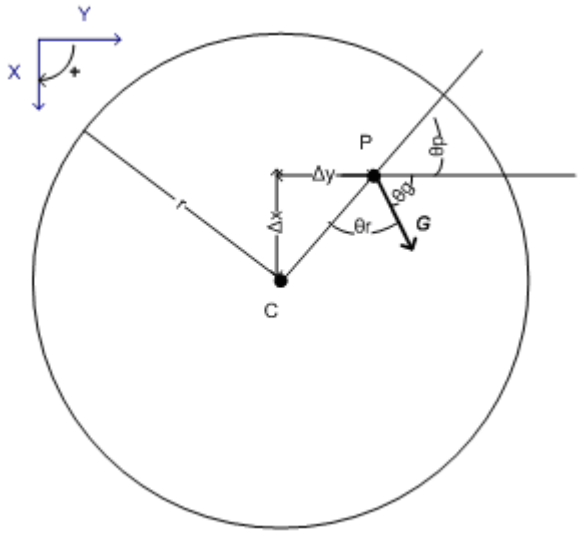


Figure 5.1. Convergence coefficient computation ($cos(\theta_r)$).

Having the values of θ_g and θ_p, (see figure 5.1) the convergence coefficient at pixel P can be yielded by

$$\cos(\theta_r) = \cos(\pi + \theta_p - \theta_g)$$

We estimated θ_p at each point as a function of its coordinates:

$$\theta_p(x, y) = \arctan\left(\frac{\Delta x}{\Delta y}\right)$$

Finally, in relation to convergence coefficient, two measures were computed:

- *Convergence coefficient mean inside:* $\mu_{\cos(\theta_r)} = \frac{1}{N}\sum_{j=1}^{N}\cos(\theta_r)_j$

- *Convergence coefficient standard deviation:*

$$\sigma_{\cos(\theta_r)} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\cos(\theta_r)_j - \mu_{\cos(\theta_r)}\right)^2}$$

Radial-gradient:

We define the radial gradient at point P (see figure 5.1) with coordinates (*x,y*), *RG(x,y)*, as

$$RG(x,y) = \left|G(x,y)\right| \times \cos(\theta_r(x,y))$$

For candidate characterisation two measures were computed:

- *Radial-gradient mean:* $\mu_{RG} = \frac{1}{N}\sum_{j=1}^{N} RG_j$

- *Radial-gradient standard deviation:* $\sigma_{RG} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(RG_j - \mu_{RG}\right)^2}$

Blob Strength:

The blob detector ties each blob candidate to a measure that quantifies the *blob strength*: $\nabla^2{}_{norm} L(x,y,\sigma)$ (see chapter 4).

5.3 Concluding Remarks

For each nodule candidate, a total of 15 features (8 intensity and contrast features and 7 features derived from first and second-order filters) were computed. They are listed in table 5.1.

Figure 5.2 shows several 2D scatter diagrams with several features to give an idea how nodule and non-nodule blobs are scattered across the feature space. Although some features are less correlated (for instance, CI versus MeanCC), it is notorious a certain degree of correlation between other couple of features, such as BS versus CBC or MeanGM versus CBC. On the other hand, for some features, the nodule blobs are densely concentrated in one region of the feature space (for instance, MeanGM).

Table 5.1 - Summary of features computed for each nodule candidate.

Intensity and Contrast features:
Intensity Mean in the Central Region (CI)
Intensity Mean along the Boundary (BI)
Intensity Maximum Value Inside (maxI)
Intensity Mean Inside (μ_I)
Intensity Standard Deviation Inside (σ_I)
Intensity Skewness Inside (s_I)
Intensity Kurtosis Inside (k_I)
Central Region - Boundary Contrast (CBC)
Features derived from first and second-order filters:
Gradient magnitude mean inside (MeanGM)
Gradient magnitude standard deviation inside (StdGM)
Convergence coefficient mean inside (MeanCC)
Convergence coefficient standard deviation inside (StdCC)
Radial-gradient mean inside (MeanRG)
Radial-gradient standard deviation inside (StdRG)
Blob Strength (BS)

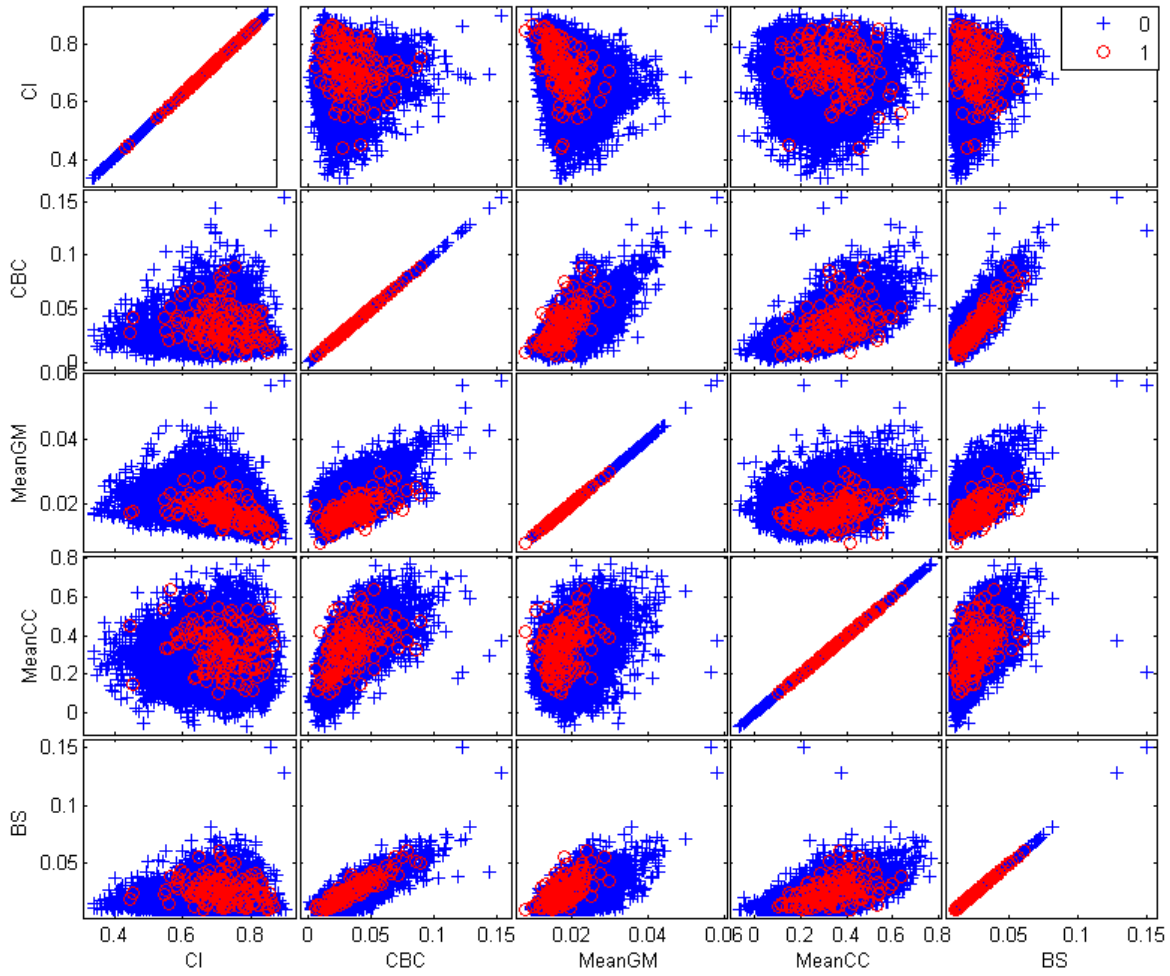


Figure 5.2 - 2D scatter diagrams (1: nodule blobs, 0: non-nodule blobs) for the following features: Intensity Mean in the Central Region (CI), Central Region - Boundary Contrast (CBC), Gradient magnitude mean inside (MeanGM), Convergence coefficient mean inside (MeanCC) and Blob Strength (BS).

Chapter 6

Classification

In this chapter, first it introduces the 1-Norm Soft Margin Support Vector Machine methodology used to classify the nodule candidates in the visible region. Then, it is described a specially conceived classification methodology for discriminating nodules from non-nodules in the hidden region.

6.1 Classification in the Visible Region

Support Vector Machines (SVMs) are a new generation of learning systems which has been employed in a wide range of applications (for instance, in the field of medical image processing, [31] – [35]). A brief overview of Support Vector Machine methodology is provided in chapter 2.

6.1.1 1-Norm Soft Margin Support Vector Machines

In the present CAD system, we employed ***1-Norm Soft Margin SVMs*** for the visible

region. In contrast to the hard margin SVMs (see chapter 2), the soft margin SVMs can tolerate noise and outliers, and take into consideration the positions of more training points than just those closest to the boundary.

Considering a set of *l* training samples, $S = \{(\mathbf{x}_1, y_1),...,(\mathbf{x}_l, y_l)\}$. We refer to the $\mathbf{x}_i$ as samples (feature vectors) and the y_i as class labels, where $y_i = \{-1,1\}$ for binary classification. The optimisation problem associated to a 1-Norm Soft Margin SVM can be summarised as follows [36]:

$$\max w(\boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

subject to:

$$\sum_{i=1}^{l} \alpha_i y_i = 0$$
$$0 \le \alpha_i \le C, i = 1,...,l$$

where C is a positive constant to ensure convergence and *K* is the kernel function. In this work, we used Gaussian kernels [36] which are given by,

$$K_{\sigma_{SVM}}(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{e}^{\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma_{SVM}^2}\right)}$$

Notice that all α_i are upper bounded by C, and then the vector $\boldsymbol{\alpha}$ is constrained to lie inside the box with length C in the positive orthant (quadrant in a two-dimensional space). Therefore, it is frequently known as *box constraint* [36]. Remarkably, the box constraints can be seen as trade-off parameters that limit the influence of outliers.

To solve the previous optimisation problem, we used the sequential minimal optimisation (SMO) algorithm (see [38]). The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets.

6.1.2 Dealing with the Imbalance Dataset Problem

Usually, the number of nodule candidates (#$class_1$) is far higher than the number of non-

nodule candidates (#class$_0$). It is a problem because the performance of SVM drops significantly when faced with imbalanced datasets where the number of negative samples far outnumbers the positive samples [40].

One strategy for dealing with imbalanced datasets (or at least attenuating this problem) is to give them different weights. Thus, the box constraints associated to non-nodule class (C_0) and to nodule class (C_1) can be set as follows:

$$C_0 = \frac{(\#class_0) + (\#class_1)}{2 \times (\#class_0)}$$

$$C_1 = \frac{(\#class_0) + (\#class_1)}{2 \times (\#class_1)} \times W$$

where W will be variable. The idea is to bias the classifier so that it pays more attention to one of the classes.

6.2 Classification in the Hidden Region

6.2.1 Why a different methodology?

For the hidden region, there were several factors that justified a different classification methodology. First of all, the total number of positive samples in this region is too small. On the other hand, during the analysis of 2-feature scatter diagrams (a typical example is shown in figure 6.2 (a)), we noticed that in some of them the nodule blobs (positive samples) were concentrated in one region of the feature space. Due to all these facts, the development of a one-class classifier to identify and fence off that region of the feature space (using the best combination of features) and then classify all candidates inside it as nodules was considered the best option.

Instead of a SVM methodology or some kind of one-class classifier derived from it, a new classifier – the *Hyper-polyhedron Classifier* – was developed, based on much

more simple tools and with a very low computational effort.

6.2.2 Hyper-polyhedron Classifier

Given a training set, this proposed one-class classifier finds a hyper-polyhedron in the feature space that includes inside or in the boundary all samples of positive class (nodule class). First, the classifier finds out the "outer points" of positive class in the feature space in order to create a convex hyper-polyhedron (a polygon in a 2-feature space or a polyhedron in a 3-feature space). To find out those "outer points", that will be the vertices of the hyper-polyhedron, an algorithm to compute the convex hull was used. The convex hull of a set of points is the smallest convex set that includes these points ([41]).

For instance, for a two-dimensional case, the convex hull algorithm chooses two points (*A* and *B*) among the given points, the leftmost and the rightmost points in the set (see figure 6.1). Then, going into the main loop, it finds the point which is furthest from the segment line *AB* (point *C*). The points which lie outside edge *AC* are put in set S_1, and points outside edge *BC* are put in set S_2 (notice that the points inside the triangle *ABC* are discarded because they cannot be on the hull). Next, the algorithm (main loop) is recursively invoked on sets S_1 and S_2.These steps are illustrated in figure 6.1. The same is done for the region below the segment line *AB* (or above, which depends the location of the first point *C*). For the set of points given in figure 6.1 (a), the final result is shown in figure 6.1 (e).

Having the classifier trained, during the classification stage, a sample will be classified as positive (nodule) if and only if it is inside the hyper-polyhedron (or over the boundary).

To illustrate how this classifier works in the feature space, figure 6.2 (b) shows the superposition of all polygons generated when we discard one different positive sample in turn. The two features used in this example were intensity kurtosis inside the inner region and central region - boundary contrast.

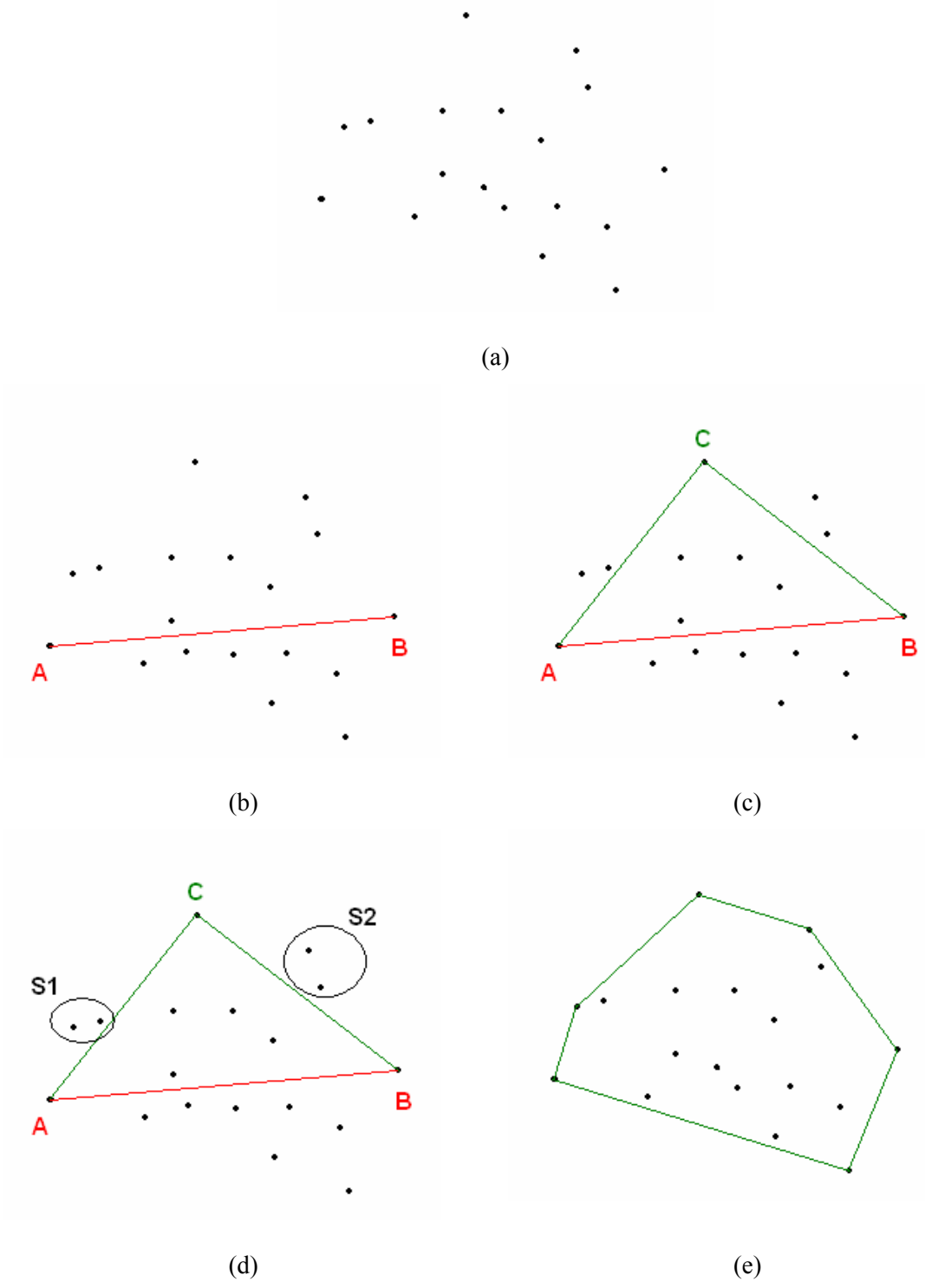
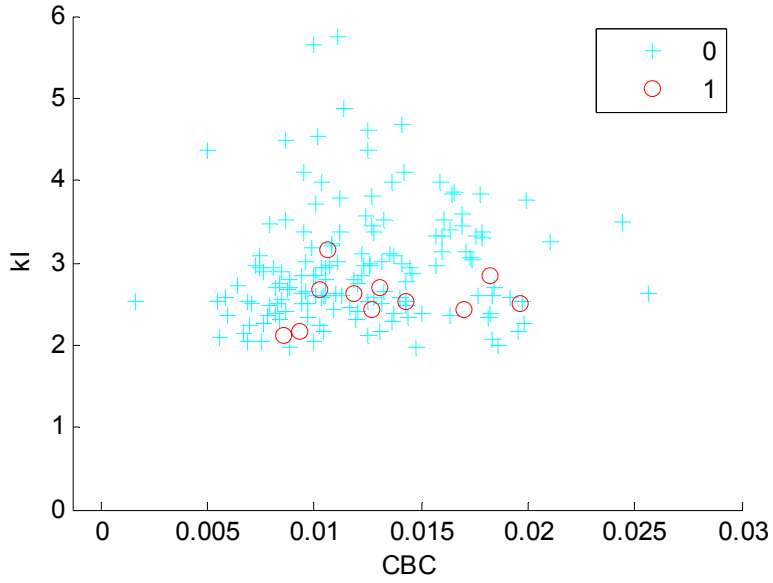
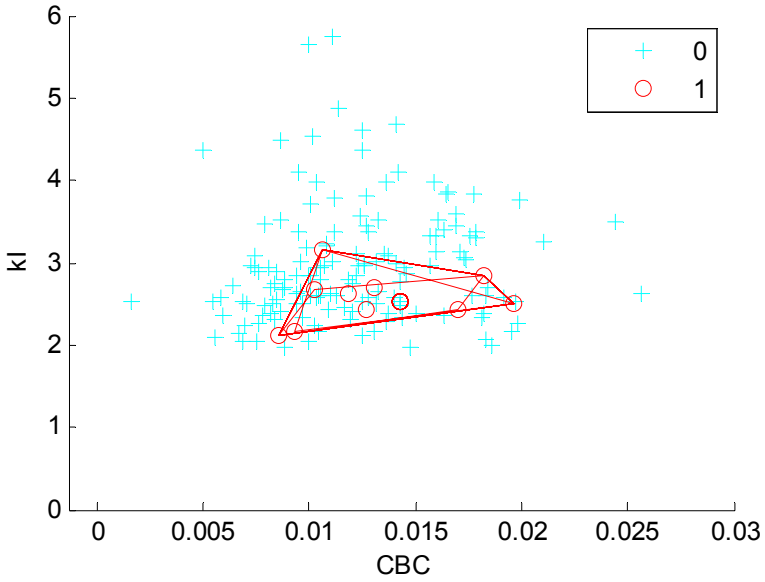


Figure 6.1 – (a) original set of point; (b, c, d) main steps of the convex hull algorithm explained in this section; (e) the convex hull of the original set of point (final result).



(a)



(b)

Figure 6.2 - (a) scatter diagram (1: nodule blobs, 0: non-nodule blobs; hidden region) for the following features: intensity kurtosis inside (kI) and central region - boundary contrast (CBC); (b) superposition of all polygons generated when we discard one different positive sample in turn.

Chapter 7

Experimental Result Analysis and Discussion

In this chapter, the experimental results for the blob detection and classification stages are reported and analysed. Finally, the overall system performance is analysed and discussed.

7.1 Nodule Candidate Detection

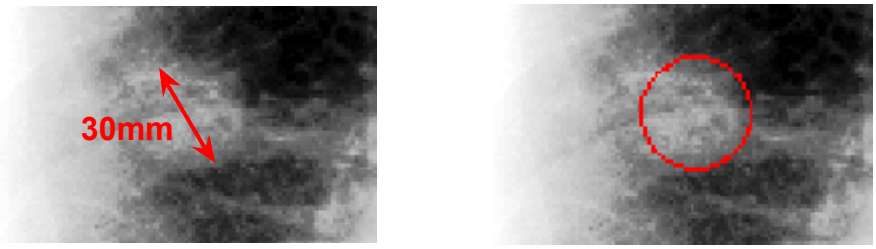
Although we had used the same blob detector in the visible and hidden regions, different values for the parameters σ and ξ were used because those regions had different characteristics, especially the contrast that was lower in the hidden region.

7.1.1 Nodule Candidate Detection – Visible Region

The first step before using the blob detector was the choice of scales and the value of

parameter ξ that would lead to good results at the blob detector output.

To analyse the effect of σ at larger scales, the blob detector was applied at scales σ = {8, 10, 12, 14}. We noticed that the biggest nodules present in the JRST database were detected at scale σ = 10, without needing to use bigger scales. Figure 7.1 shows one case whose nodule is one of the biggest ones and it was detected at scale σ =10.



(a) (b)

Figure 7.1 - (a) Nodule of the case JPCLN017; (b) Response of the blob detector at scale σ=10.

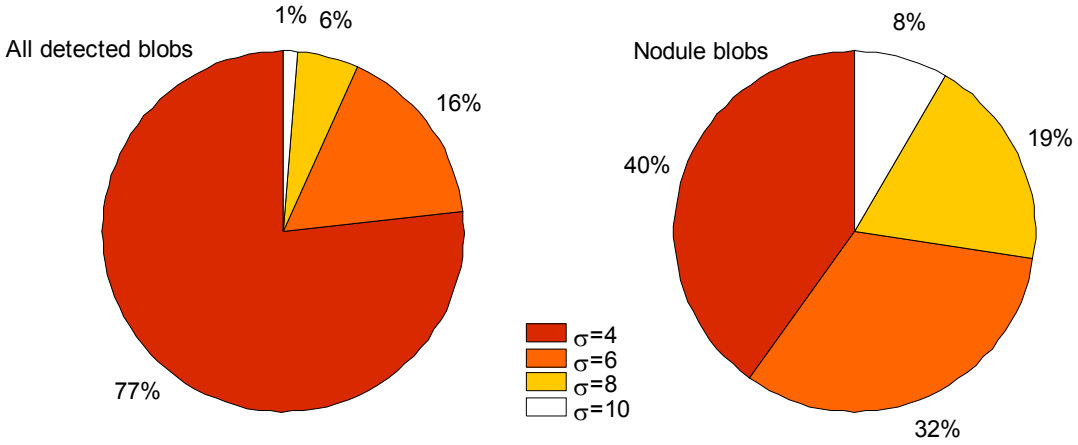
As far as the choice of the smallest scale is concerned, at first, the scale σ = 2 was included. However, it leaded to a huge number of false positives (185.9 in average) and the experiment was repeated without σ = 2. The average number of false positives was reduced to 64.6. The results are summarized in table 7.1 (the reason why ξ=0.01 is explained later).

To achieve a trade-off between the number of detected nodules and the number of false positive, we decided to discard scale σ =2 and use only scales σ = {4, 6, 8, 10 }.

End-on vessels are one of the main responsible for the exponential increase in detected blobs at smaller scales. The incident X-ray beam, parallel to the vessel, creates small round spots which contain high contrast and a high degree of circularity. Thus, it is very difficult to distinguish between subtle nodules and end-on vessels. In figure 7.2 is illustrated the proportion of blobs at different scales σ, considering all the blobs (detected by the blob detector) and considering only the blobs that are actually nodules. As expected, the majority of blobs (at the output of the blob detector) belong to scale σ =4.

Table 7.1 - Blob detector performance with and without scale σ = 2. All the 141 radiographs with a lung nodule in the visible region were used.

	σ = { 2, 4, 6, 8, 10 } ξ = 0.01	σ = { 4, 6, 8, 10 } ξ = 0.01
Detected nodules (out of 141)	136	134
FP per image (average)	185.9	64.6



(a) (b)

Figure 7.2 - Proportion of blobs detected at different scales σ, considering all blobs (a), and considering only blobs that are actually nodules (b).

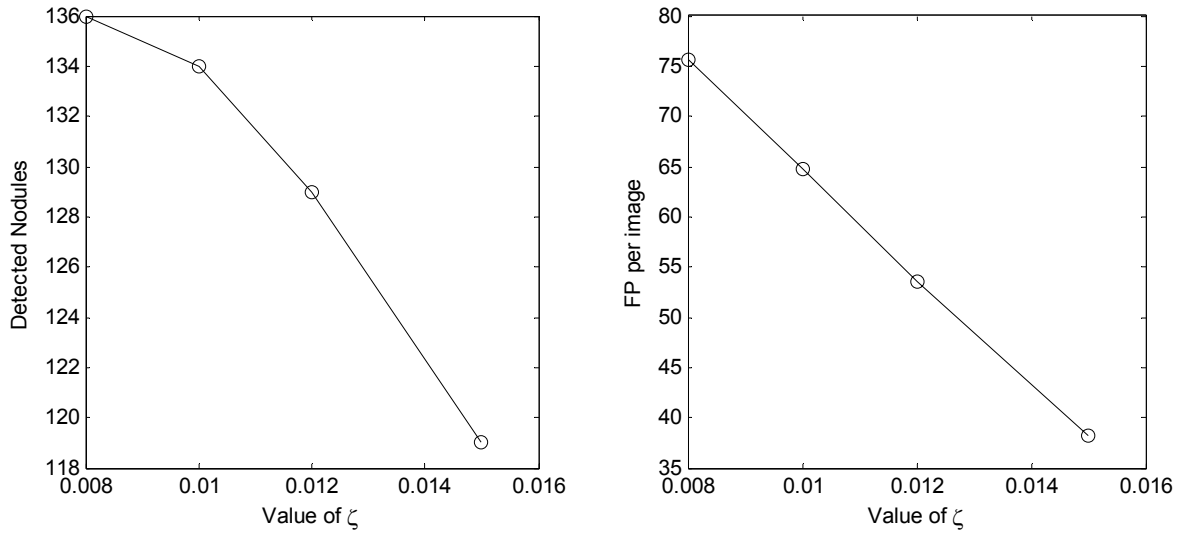
Another important parameter is the value of ξ. In fact, it makes the blob detector more or less sensitive. In order to find out the best one, different values of ξ were tested. The results are reported in table 7.2 and are graphically illustrated in figure 7.3 (for an example, see figure 4.4)

To achieve a trade-off between the number of detected nodules and the number of

false positive, we decided to use ξ =0.01. In fact, a decrease of ξ to 0.008 leads to an increase by 1.5% in the detected nodules together with a very disadvantageous increase by 18.5% in the FPs per image.

Table 7.2 - Blob detector performance for several values of ξ and σ = {4, 6, 8, 10}. All the 141 radiographs with a lung nodule in the visible region were used.

	Value of ξ			
	0.008	0.01	0.012	0.015
Detected nodules (out of 141)	136	134	129	119
FP per image (average)	76.7	64.6	53.5	38.3



(a) (b)

Figure 7.3 - Detected nodules (a) and FP per image (b) as a function of parameter ξ (σ = {4, 6, 8, 10}).

After the candidate detection stage in the visible region, there were 134 nodule blob

samples and 9949 non-nodule blob samples from patients with a nodule (154 images). Two typical examples of blob detector outputs in this region are shown in figure 7.4 and figure 7.5. The radiograph in figure 7.5 was zoomed in on the nodule region (see figure 7.6).

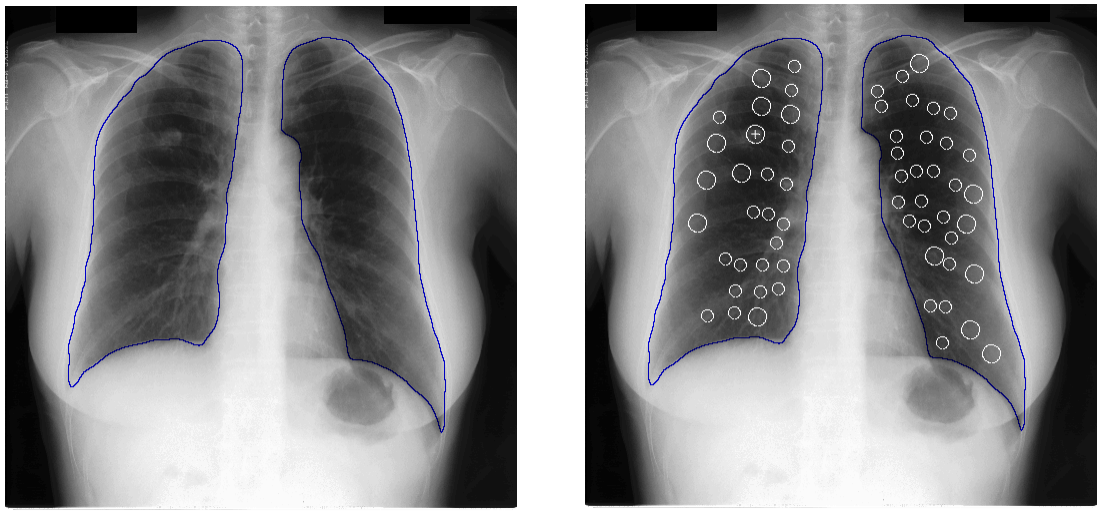


Figure 7.4 - Case JPCLN006: nodule candidates selected by the multi-scale blob detector in the visible region with $\sigma = \{ 4, 6, 8, 10 \}$ and $\xi=0.01$ (+: true nodule).

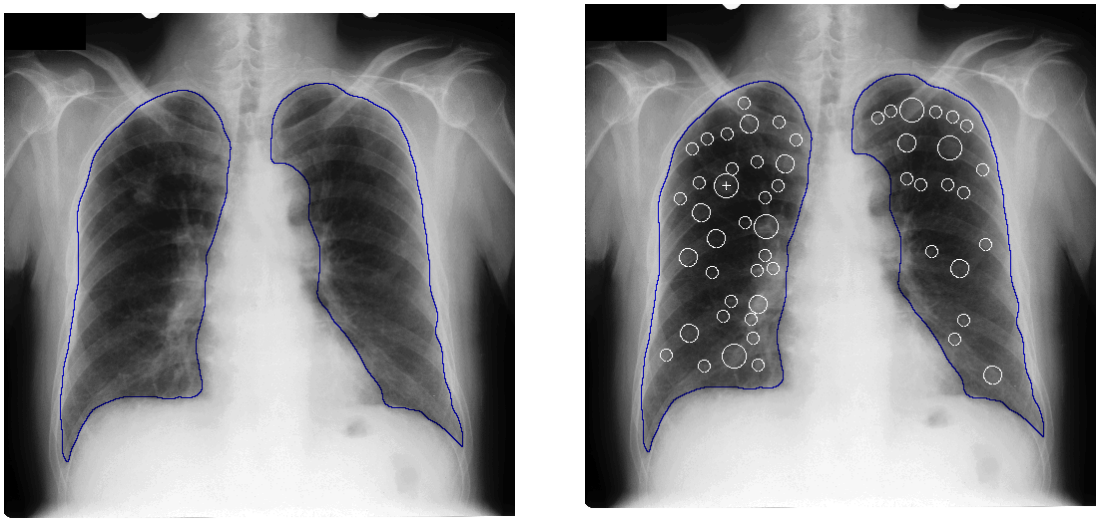
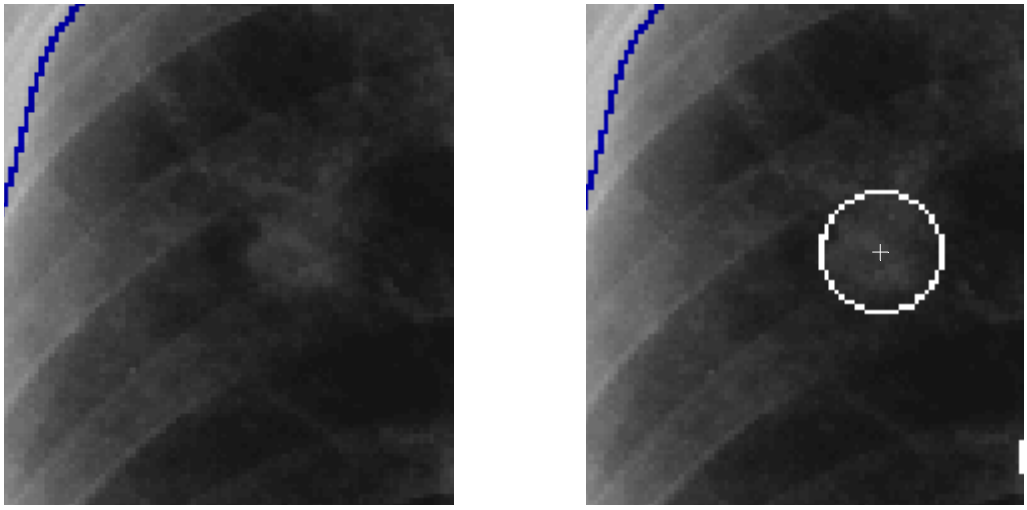


Figure 7.5 - Case JPCLN014: nodule candidates selected by the multi-scale blob detector in the visible region with $\sigma = \{ 4, 6, 8, 10 \}$ and $\xi=0.01$ (+: true nodule).



(a) (b)

Figure 7.6 - Case JPCLN014: (a) it was zoomed in on the nodule region; (b) the blob detected over the true nodule (+).

7.1.2 Nodule Candidate Detection – Hidden Region

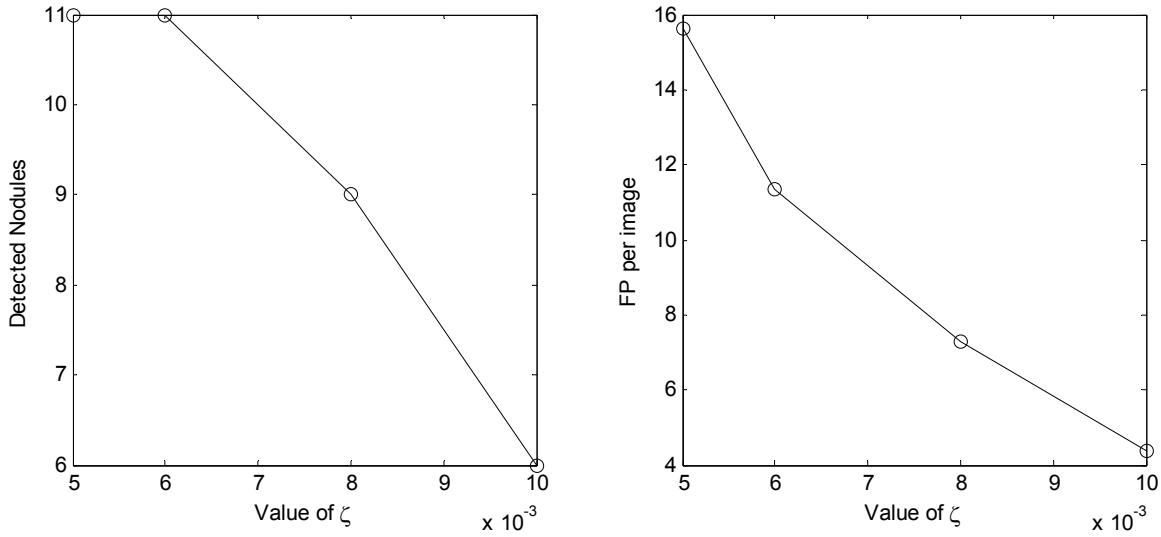
In this region, the x-ray images have a low contrast and consequently the task of blob detection requires more sensitivity. We used scales $\sigma = \{4, 8\}$ because the inclusion of scales $\sigma = \{2, 10\}$ did not increase the number of detected blobs in the hidden region. The results for different values of ξ are reported in table 7.3 and are graphically illustrated in figure 7.7. For hidden regions, we decided to use $\xi = 0.006$ which leads to 11 out of 13 nodules and at the same time with relatively few false positives per image.

After the candidate detection stage in the hidden region and considering the 13 patients with a nodule in this region, there were 11 nodule blob samples and 147 non-nodule blob samples.

A typical example of the blob detector output in this region is shown in figure 7.8. The radiograph was zoomed in on the nodule region (see figure 7.9).

Table 7.3 - Blob detector performance for several values of ξ and $\sigma = \{4, 8\}$. All the 13 radiographs with a lung nodule in the hidden region were used.

	Value of ξ			
	0.005	0.006	0.008	0.01
Detected nodules (out of 13)	11	11	9	6
FP per image (average)	15.6	11.3	7.3	4.4



(a) (b)

Figure 7.7 - Detected nodules (a) and FP per image (b) as a function of parameter ξ. ($\sigma = \{4, 8\}$)

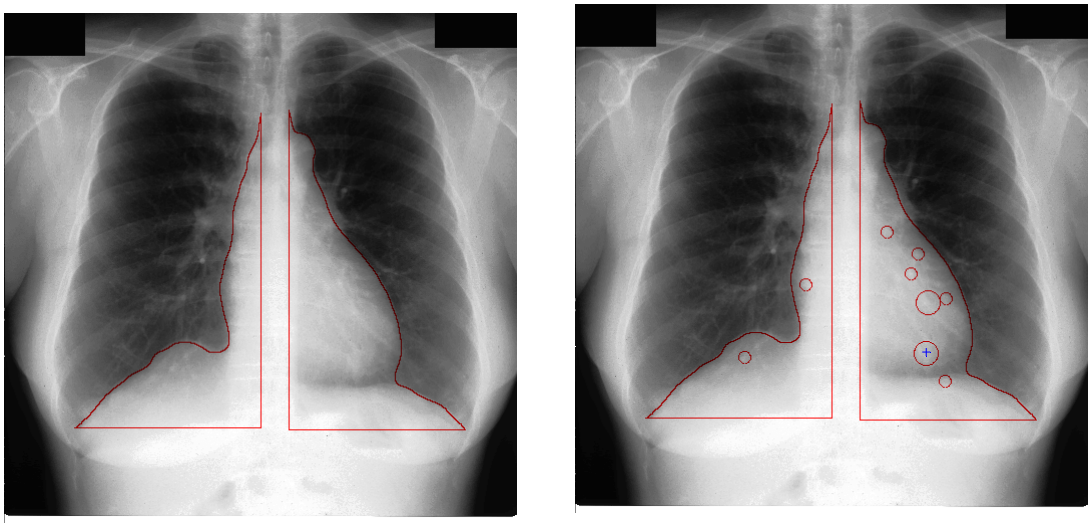
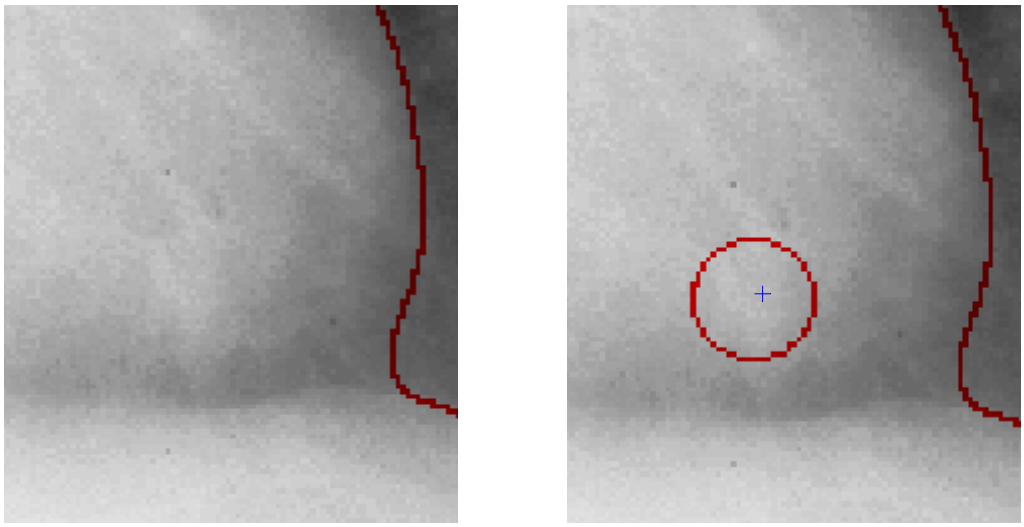


Figure 7.8 - Case JPCLN113: nodule candidates selected by the multi-scale blob detector in the hidden region with $\sigma = \{ 4, 8 \}$ and ξ=0.006 (+: true nodule).



(a) (b)

Figure 7.9 - Case JPCLN113: (a) it was zoomed in on the nodule region; (b) the blob detected over the true nodule (+).

7.2 Classification in the Visible Region

After the candidate detection in the visible region, there were 134 nodule blob samples

and 9949 non-nodule blob samples from patients with a nodule (154 images). We decided to add 551 non-nodule blob samples from normal cases, increasing to 10500 the total number of non-nodule blob samples. The proportion of nodule candidates to non-nodule candidates was $\approx 1 / 78$. Using the strategy proposed in chapter 6, the box constraints associated to non-nodule class (C_0) and to nodule class (C_1) were set as follows:

$$C_0 = \frac{(\#class_0) + (\#class_1)}{2 \times (\#class_0)} \approx 0.506$$

$$C_1 = \frac{(\#class_0) + (\#class_1)}{2 \times (\#class_1)} \times W \approx 39.679 \times W$$

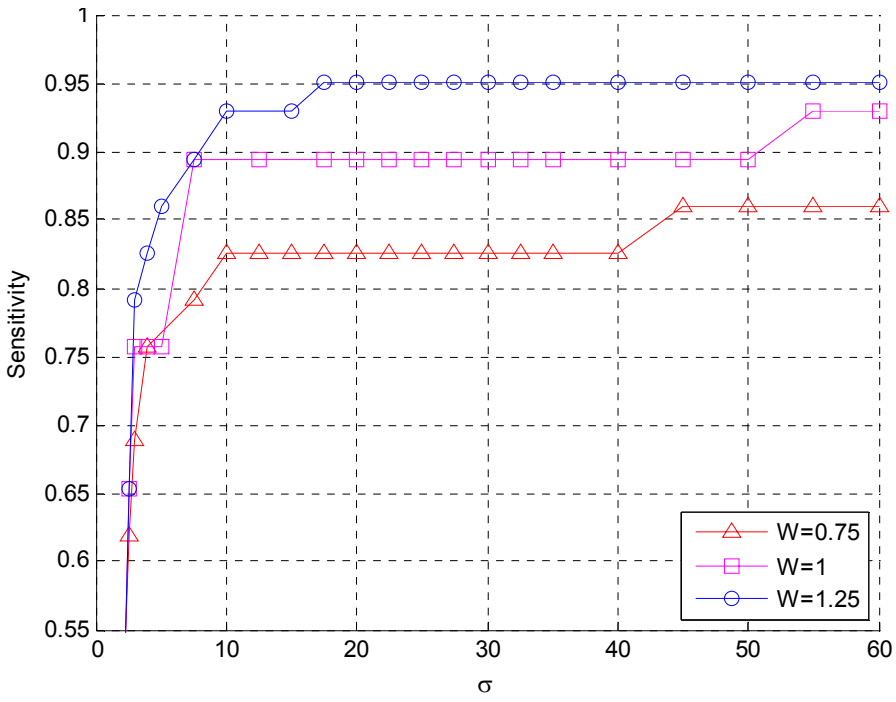
During the experiments, we used three different values for W: 0.75; 1; 1.25. For each value of W, we tested the following values of σ_{SVM}: 1; 2; 2.5; 3; 4; from 5 to 60 with equal spaces of 2.5; 80; 100. In total, 90 different SVMs were tested.

The validation methodology used here was 5-fold cross-validation. The images of the JSRT database were split over five sets, with images of the same subtlety class distributed evenly over the sets. One set is then classified with a classifier previously trained on the four remaining sets. This is repeated five times to classify all samples. The final results, for each classifier, were averaged over these five trials.

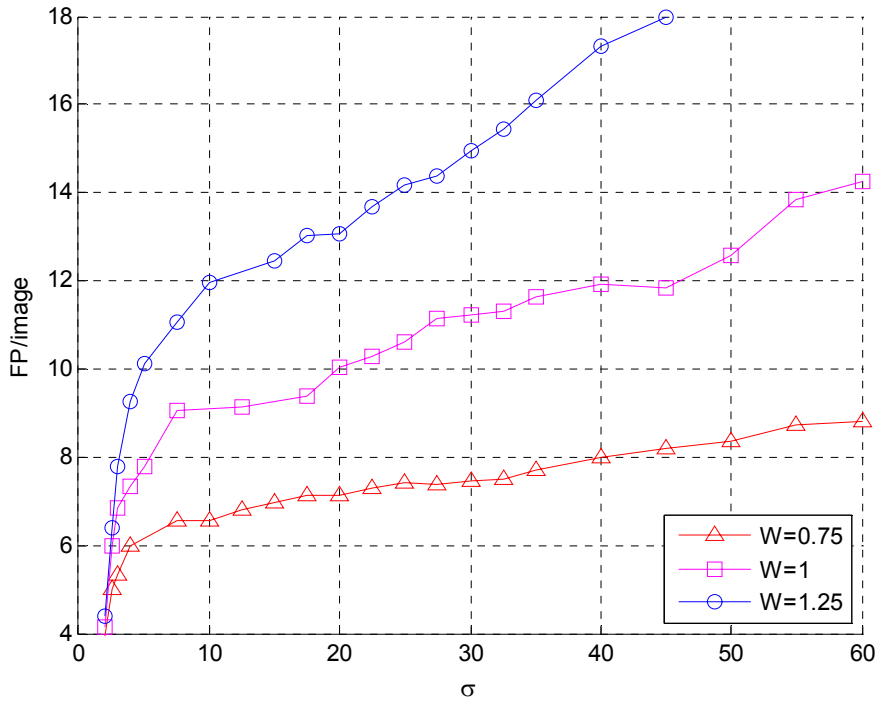
Figure 7.10 illustrates how sensitivity and FPs per image vary when the parameter σ_{SVM} increases. At least for values of σ_{SVM} between 1 and 60, an increase in the value of σ_{SVM} leads to a greater sensitivity but leads to a greater number of FPs per image as well.

In order to choose the operating point, the sensitivity versus the average number of false positives per image was plotted and is shown in figure 7.11. It should be noticed that only 134 out of 141 nodules were detected in the visible region by the blob detector. Hence, the maximum achievable value of sensitivity by the CAD system in this region is 95.04%.

At 6.5 FPs per image, the sensitivity of the system in the visible region is 83% when W=0.75 and σ_{SVM} = 10. *W*hen sensitivity increases to 86%, the FPs per image increase to about 8 (W = 0.75 and σ_{SVM} = 45). When W = 1 and σ_{SVM} = 7.5, a sensitivity of about 90% is achieved with 9 FPs per image, which is not very high at this sensitivity.



(a)



(b)

Figure 7.10 - Sensitivity (a) and FPs per image (b) as a function of parameter σ_{SVM}.

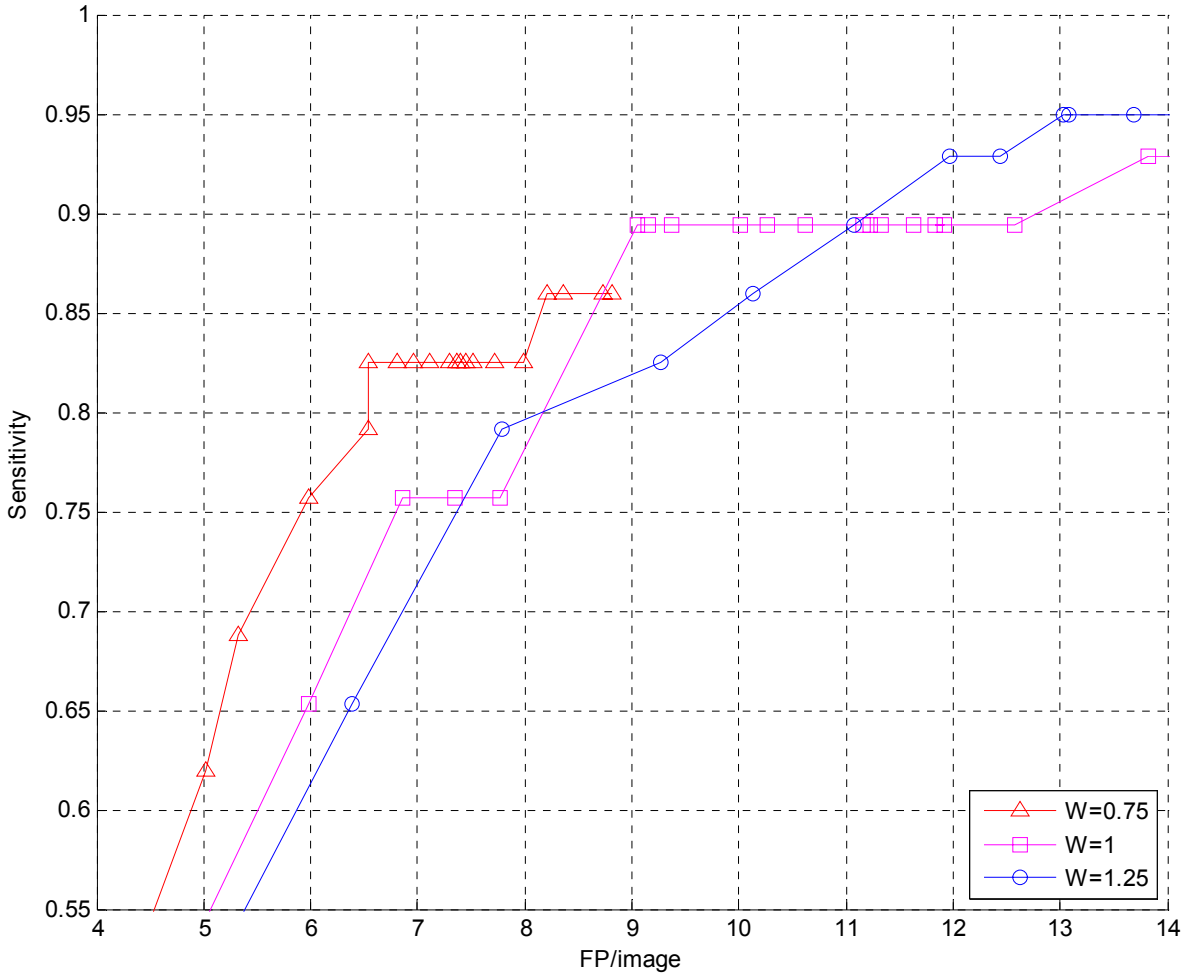
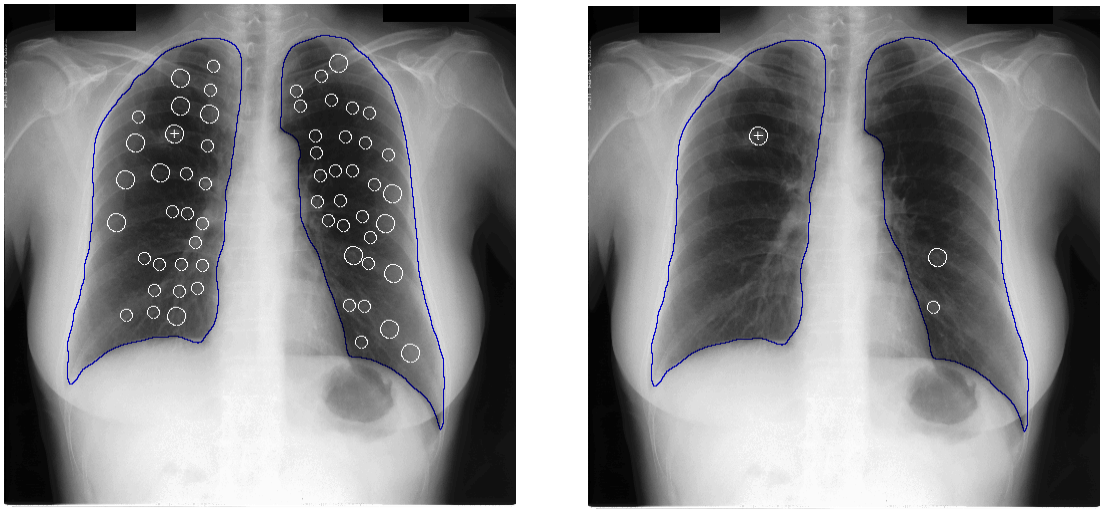


Figure 7.11 - Sensitivity as a function of FPs per image for several Gaussian SVMs.

Lastly, it was important to choose the parameters W and σ_{SVM} that would be used in the CAD system, i.e., the operating point. At the same time, we want to have as greater sensitivity as possible and as less FPs per image as possible. Attaching more weight to the latter and taking into account the graphs in figure 7.11, the curve of W=0.75 is the best one to serve this purpose, especially the knee of this curve. Thus, the point with sensitivity equals 83% and 6.5 FPs per image was chosen as the operating point for the SVM (W = 0.75 and σ_{SVM} = 10).

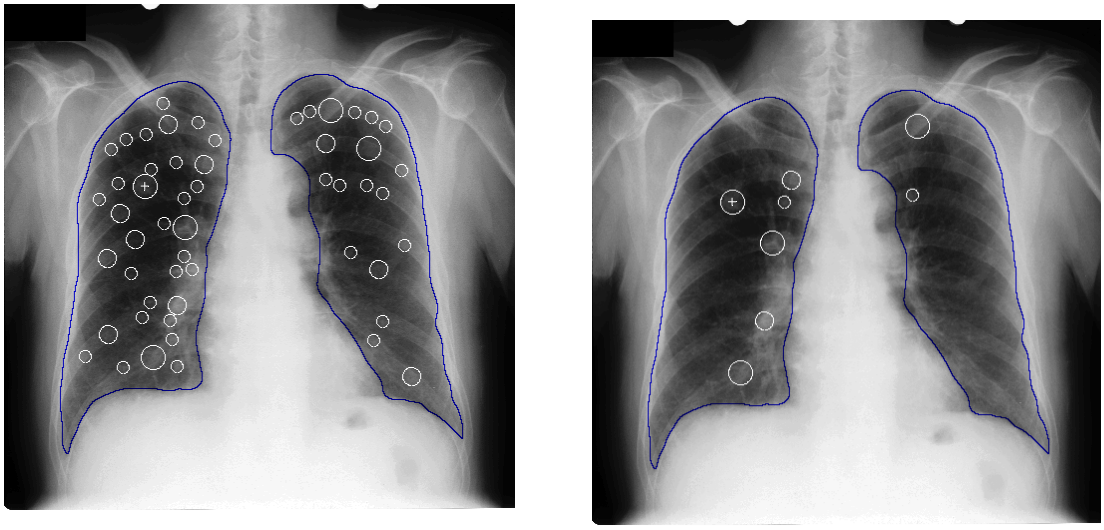
The following figures show two output images for this classifier. In the first case (figure 7.12), the number of FPs per image obtained (3) was less than the average (6.5) whereas in the second case (figure 7.13), the number of FPs per image (8) was greater

than the average. It should be noticed that there are some anatomic structures that misled the classifier. These observations will be discussed later.



(a) (b)

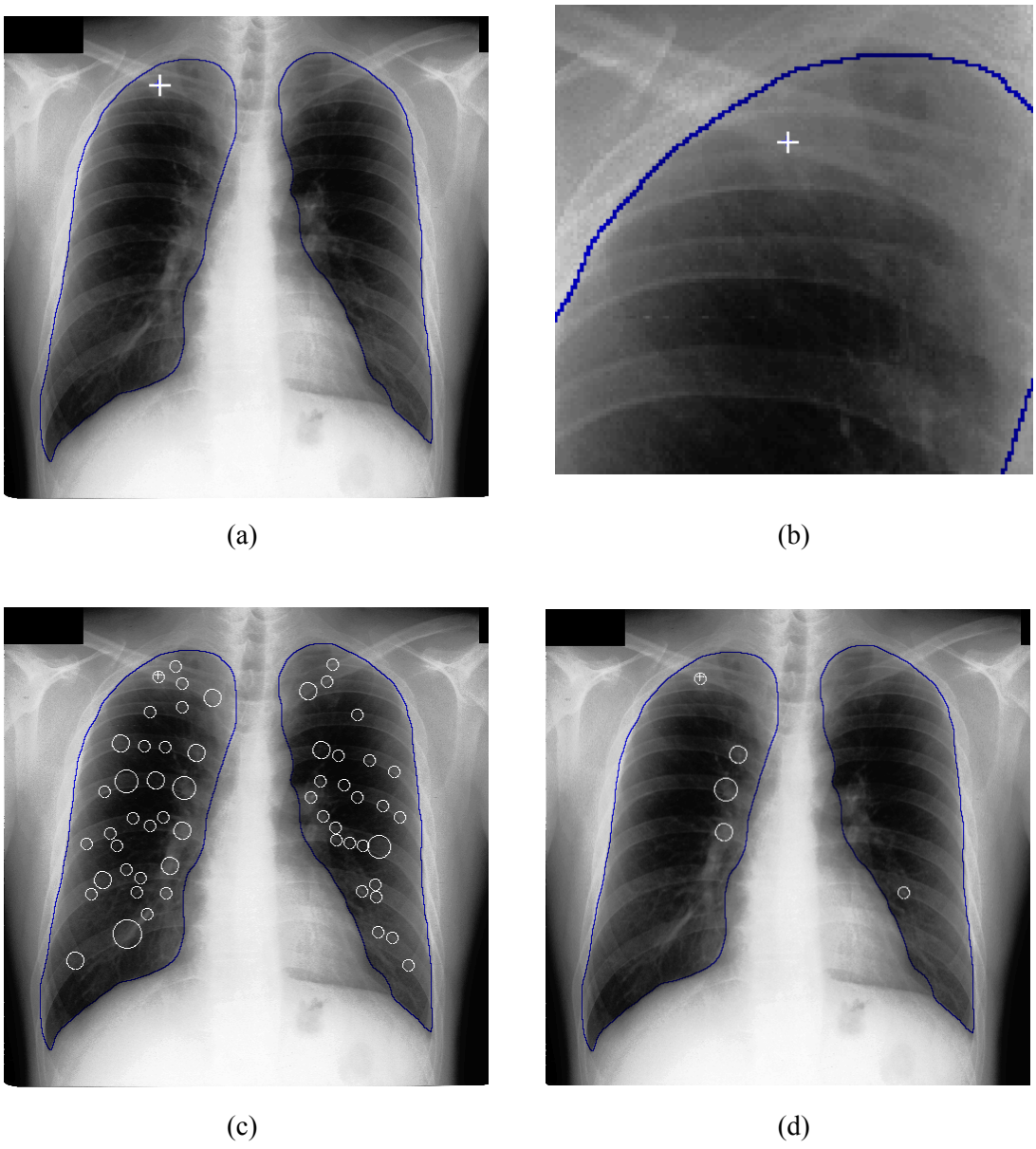
Figure 7.12 - Case JPCLN006: nodule candidates before (a) and after (b) the classification stage. The Gaussian SVM with $W = 0.75$ and $\sigma_{SVM} = 10$ was used (+: true nodule).



(a) (b)

Figure 7.13 - Case JPCLN014: nodule candidates before (a) and after (b) the classification stage. The Gaussian SVM with $W = 0.75$ and $\sigma_{SVM} = 10$ was used (+: true nodule).

Figure 7.14 presents another case with a small nodule that lies in the edge of the clavicle. This nodule was graded by radiologist as extremely subtle [4]. Figure 7.14 (d) shows the nodule candidates before and after the classification stage where the real nodule was correctly detected and classified by the CAD system.



(a) (b)

(c) (d)

Figure 7.14 – Case JPCLN139: (a) original radiograph; (b) radiograph was zoomed in on the nodule region; (c) after the blob detector; (d) after the classification stage. The Gaussian SVM with W = 0.75 and σ_{SVM} = 10 was used (+: true nodule).

7.3 Classification in the Hidden Region

In this region the Hyper-polyhedron Classifier was employed (see chapter 6). At first, we decide to use 2 features. As we did not know what the optimal 2-feature combination was, for each one (105 in total), we carried out the following method:

One image from the 13 images with a nodule in the hidden region was held out of the training set (we left one out to improve the generalization). Then, we tested the classifier on all images. This process was repeated 13 times with all different training sets. The final results for each 2-feature subset were averaged over these 13 trials.

The final results using all 2-feature subsets are summarised in the form of box plots in figure 7.15 where we obtained an averaged sensitivity of about 81.4% and, on average, about 4 FPs per image. The best results are reported in table 7.4.

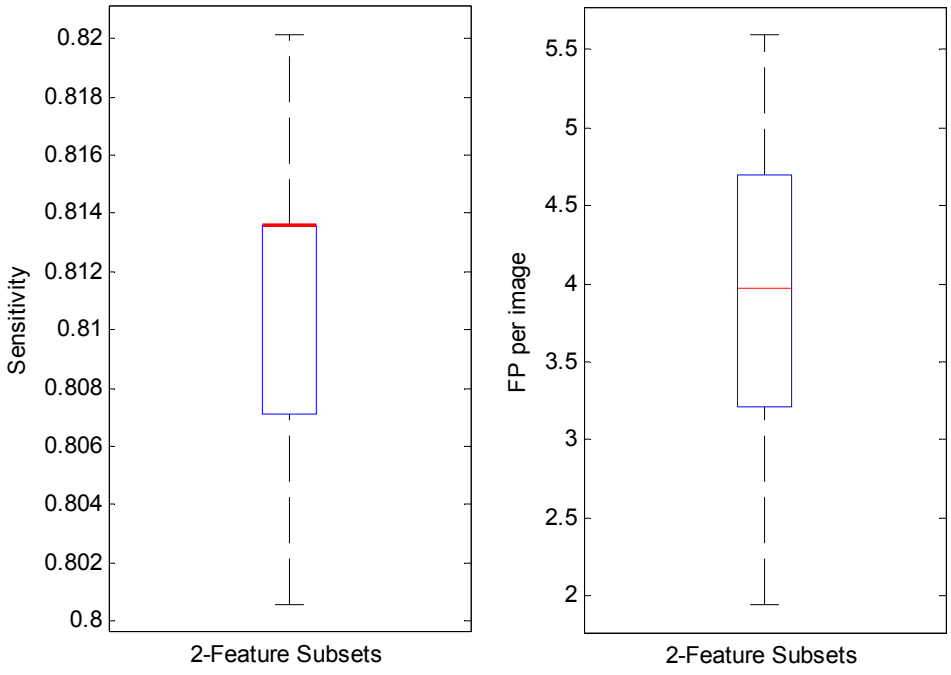


Figure 7.15 – Results obtained with the Hyper-polyedron Classifier in the hidden region using 2-feature subsets.

Table 7.4 - Best results obtained with the Hyper-polyedron Classifier in the hidden region using 2-feature subsets.

Criterion	Best 2-feature subset	Sensitivity	FP per image
max (Sensitivity)	{ k_I ; CBC }	0.8201	2.725
min (FP per image)	{ MeanGM ; StdGM}	0.8071	1.947

In order to decrease the number of false positives, we carried out the previous experiments but now with 3-feature subsets. Again, as we did not know what the optimal 3-feature subset was, we repeated the previously explained method for all possible 3-feature subsets (455 in total).

The final results using 3-feature subsets are summarised in the form of box plots in figure 7.16 where we obtained an averaged sensitivity of about 80.7% and, on average, about 3.2 FPs per image. The best results are reported in table 5.1.

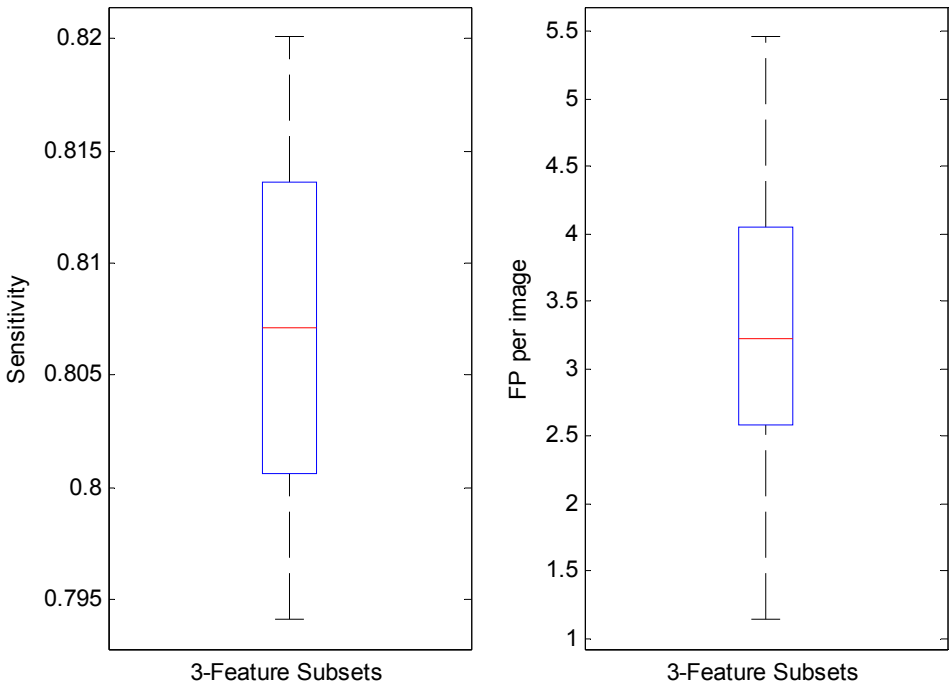
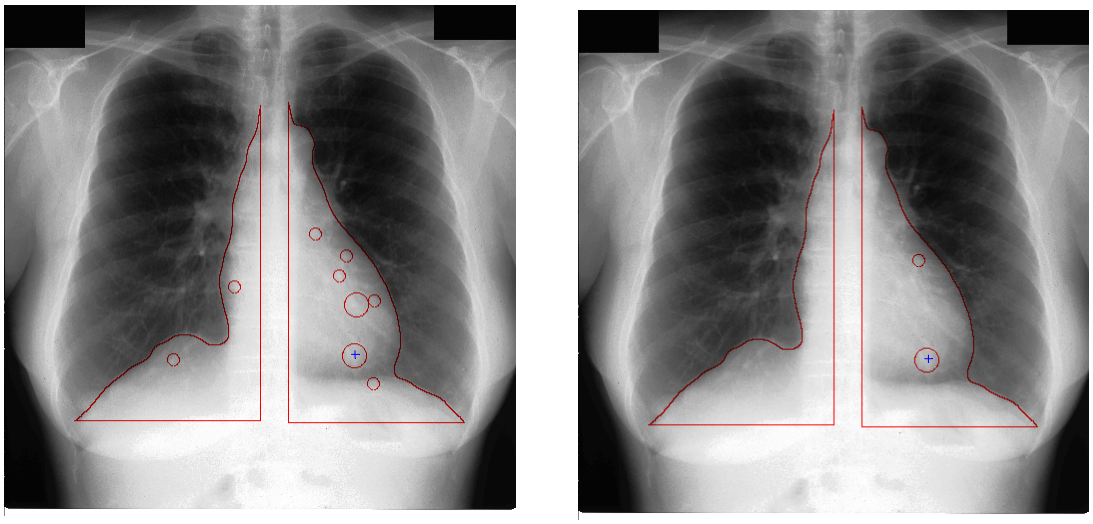


Figure 7.16 – Results obtained with the Hyper-polyedron Classifier in the hidden region using 3-feature subsets.

With the features: blob strength (BS), intensity kurtosis (k_I) and convergence coefficient mean (MeanCC) inside the inner region, a sensitivity of 82% with only about 1.8 FPs per image was reached (table 5.1). A typical example of the classifier output using these features is shown in figure 7.17.

Table 7.5 - Best results obtained with the Hyper-polyedron Classifier in the hidden region using 3-feature subsets.

Criterion	Best 3-feature subset	Sensitivity	FP per image
max (Sensitivity)	**{ kI ; MeanCC ; BS }**	**0.8201**	**1.838**
min (FP per mage)	{ sI ; MeanGM ; BS }	0.8071	1.137



(a) (b)

Figure 7.17 - Case JPCLN113: nodule candidates before (a) and after (b) the classification stage. The following features were used: intensity kurtosis inside (kI), convergence coefficient mean inside (MeanCC) and blob strength (BS) (+: true nodule).

7.4 Overall System Performance

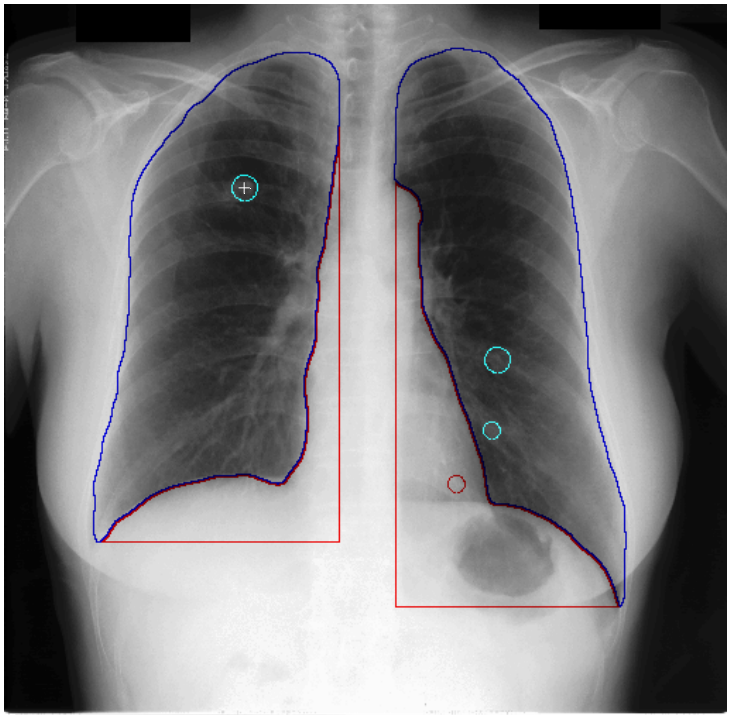
In the last step, the combination of the two sub-systems (for the visible and hidden regions) is done simply by overlapping the candidates classified as nodules by each sub-system. The overall performance (using the operating points previously chosen) can be characterised by an overall sensitivity of:

$$S_T = \frac{141}{154} \times 0.83 + \frac{13}{154} \times 0.82 \approx 0.83$$

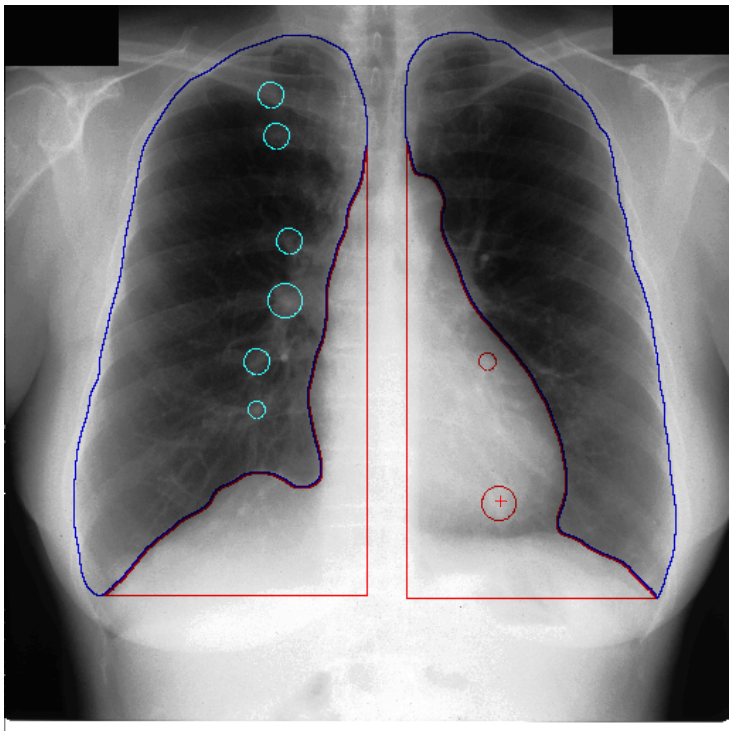
Two typical CAD outputs are shown in figure 7.18. In figure 7.18 (a) the patient had a nodule in the visible region and in figure 7.18 (b) the patient had a nodule in the hidden region. In both cases the nodule was detected. However, some anatomic structures misled the classifier.

Careful observation was conducted to understand why several candidates were falsely classified as nodules. As observed, many false positives were generated by overlapping bony structures, especially ribs with other ribs, clavicle or scapula. Moreover, several false positives were generated by end-on vessels. In fact, the incident X-ray beam, parallel to the vessel, creates small round spots which contain high contrast and a high degree of circularity. Thus, it is very difficult to distinguish between subtle nodules and end-on vessels. Figure 7.19 shows two typical examples of false positives.

One the other hand, there are nodules that were not detected by the present CAD system. The majority of those nodules were among the very subtle or extremely subtle cases. Those nodules were very small and dim and some of them were located close to the lung edge. An illustrative example is the radiograph present in figure 7.20 with a very small nodule located close to the lung edge. This nodule was not detected in the blob detector stage due to its dimness.



(a)



(b)

Figure 7.18 - Typical CAD outputs: (a) case JPCLN006; (b) case JPCLN113. (+: true nodule).

(a) (b)

Figure 7.19 - Typical examples of false positives: (a) overlapping ribs; (b) end-on vessel.

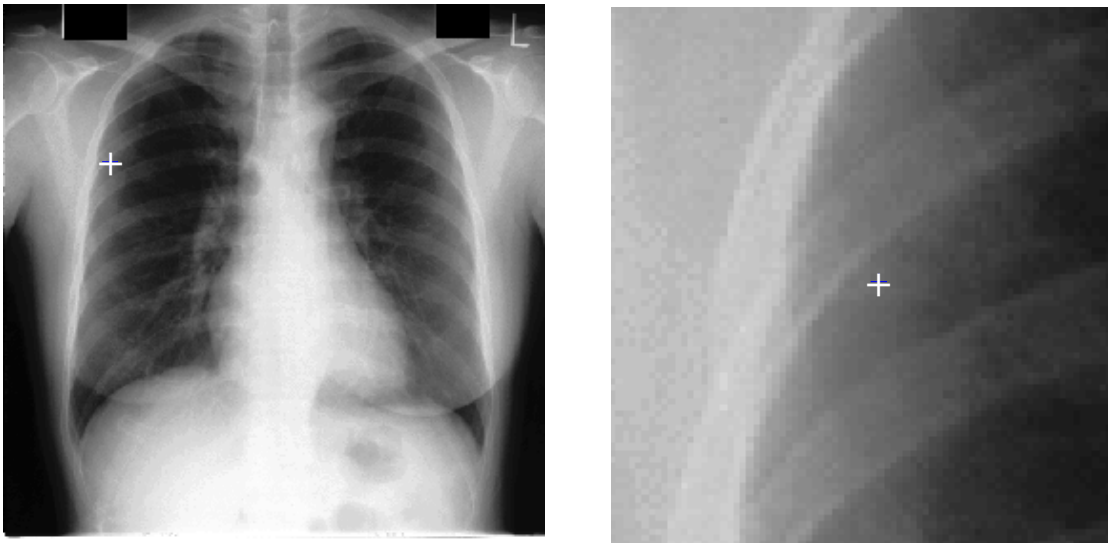


Figure 7.20 - Case JPCLN145 with a very small nodule located close to the lung edge. This nodule was not detected in the blob detector stage due to its dimness (+: true nodule).

7.5 Concluding Remarks

Considering only the visible region, the results appear to be promising when compared with other CAD systems. For instance, Ginneken et al. [8], who used a similar blob detector, obtained after the classification stage (two KNN classifiers) a sensitivity of

about 70% at 6.5 FPs per image and about 74% at 9 FPs per image. They only considered nodules in the visible region. Scaling the sensitivity that we obtained with the present CAD system by a factor of 141/154 = 0.9156, in order to use the 154 nodules in the JRST database as the reference for 100%, we obtained better results: 76% at 6.5 FPs per image and 82.4% at 9 FPs per image.

However, a straightforward comparison with other CAD systems found in the literature (for instance, [14], [29]) is very difficult due to the variability in the training and testing datasets, the use of different labelling methods and also the use of different validation methodologies, which lead to inaccurate conclusions. It also should be noticed that many CAD systems consider a candidate as a TP without requiring that the candidate region (suspicious region) must include the centre of the true nodule. Consequently, it leads to overrated results when compared to CAD systems, such as the present system, where that restriction was imposed.

Moreover, the present CAD system also considers nodules in the hidden region. It adds value to the CAD system, boosting its performance. In fact, the majority of CAD systems presented in the literature excludes these nodules, although they represent about 8.5% of the total number of nodules in the JRST database.

Chapter 8

Conclusion

A new CAD system for lung nodule detection in chest radiographs has been presented. Based on the results reported in the previous chapters, we believe that this system offers some useful innovations and ideas for future research projects.

In fact, a novel combination of a multi-scale blob detector and a Support Vector Machine was used for the visible region where good results were obtained: a sensitivity of 83% at 6.5 false positives (FP) per image and 90% at 9 FP per image. Moreover, with the combination of the multi-scale blob detector and the *Hyper-polyhedron Classifier* in the hidden region, a sensitivity of 82% at 1.8 FP per image was reached.

The block diagram of the complete CAD system with the methodologies chosen at each stage is shown in figure 8.1.

8.1 Future Research

In future research, a possible strategy to reduce the number of false positives can make use of information about patient's anatomic structures. For instance, the automatic

detection of rib borders in chest radiographs proposed by Moreira et al. [12] could be used to assess the rib regions using extra tools and understand if a candidate is really a nodule or not. Notice that the intersection between ribs or ribs and clavicle is a rough lozenge and not a rough circle. We believe that this kind of strategies would reduce the false positives in these regions.

On the other hand, the use of the original images (2048×2048) instead of the down-sampled images (512×512) might ease the detection of the smallest nodules.

Furthermore, in future research projects, it would be useful the introduction of an algorithm to determine whether a given nodule is benign or malign, for example, based on candidate boundary shape.

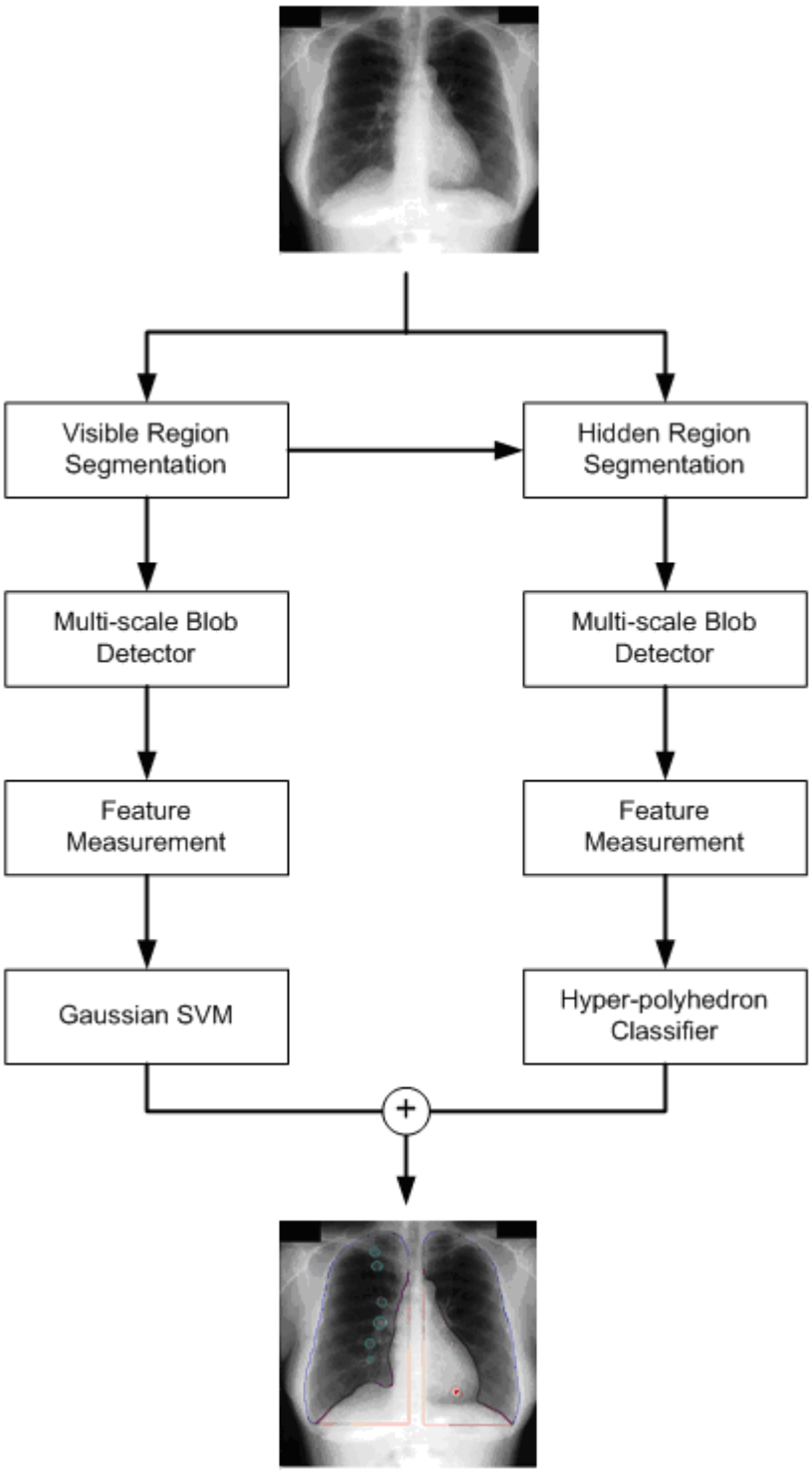


Figure 8.1 - The block diagram of the complete CAD system with the methodologies chosen at each stage.

References

[1] World Health Organization, "Cancer - Fact sheet N°297," Feb. 2006. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/index.html. [Accessed: Jan. 14, 2008].

[2] Radiological Society of North America, "Radiation Exposure in X-ray Examinations," Feb. 2007. [Online]. Available: http://www.radiologyinfo.org/en/safety/index.cfm?pg=sfty_xray&bhcp=1. [Accessed: Jan. 16, 2008].

[3] B. Keserci and H. Yoshida, "Computerized detection of pulmonary nodules in chest radiographs based on morphological features and wavelet snake model," *Medical Image Analysis*, vol. 6, pp. 431-447, 2002.

[4] 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," *Amer. J. Roentgenol*., vol. 174, pp. 71–74, 2000.

[5] B. van Ginneken. "Computer-aided Diagnosis in Chest Radiography," Ph.D. dissertation, Utrecht University, The Netherlands, March 2001.

[6] R. C. Gonzalez and R. E. Woods, Digital Image Processing, second edition. Upper Saddle River: Prentice Hall, 2002.

[7] R. C. Gonzalez, R. E. Woods and S. L. Eddins, Digital Image Processing using Matlab. Upper Saddle River: Pearson Education, 2004.

[8] A. Schilham, B. van Ginneken and M. Loog, "A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database," *Medical Image Analysis*, vol. 10, pp. 247-258, 2006.

[9] H. Zhao, S. Lo, M. Freedman, and S. Mun, "On automatic temporal subtraction of chest radiographs and its enhancement for lung cancers," *Proc. SPIE*, vol. 4322, pp. 1867–1872, 2001.

[10] Q. Li, S. Katsuragawa, T. Ishida, H. Yoshida, S. Tsukuda, H. MacMahon, and K. Doi, "Contralateral subtraction: A novel technique for detection of asymmetric abnormalities on digital chest radiographs," *Med. Phys.*, vol. 27, no. 1, pp. 47–55, 2000.

[11] Q. Li, S. Katsuragawa, and K. Doi, "Improved contralateral subtraction images by use of elastic matching technique," *Medical Physics*, vol. 27, no. 8, pp. 1934–1942, 2000.

[12] R. Moreira, A. M. Mendonça and A. Campilho, "Detection of Rib Borders on X-ray Chest Radiographs," *Lecture Notes in Computer Science*, vol. 3212, pp. 108-115, 2004.

[13] B. Ginneken, B. Romeny, "Automatic segmentation of lung fields in chest radiographs", Medical Physics, 27, pp. 2445-2455, 2000.

[14] R. C. Hardie, S. K. Rogers, T. Wilson and A. Rogers, "Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs," *Medical Image Analysis* (in press), 2008.

[15] A. M. Mendonça, J. A. Silva and A. Campilho, "Automatic delimitation of lung fields on chest radiographs," *IEEE International Symposium on Biomedical Imaging*, vol. 2, pp. 1287- 1290, April 2004.

[16] M. Monteiro, A. M. Mendonça and A. Campilho, "Lung Field Detection on Chest Radiographs", *Proceedings of VIIP02, IASTED International Conference on Visualization, Imaging and Image Processing*, J. Villanueva (eds.), Spain, pp. 367-372, 2002.

[17] M. J. Carreira, D. Cabello, M. G. Penedo and A. Mosquera, "Computer-aided diagnoses: automatic detection of lung nodules," *Medical Physics*, vol. 25, no.10, pp. 1998–2006, 1998.

[18] B. Keserci and H. Yoshida, "Computerized detection of pulmonary nodules in chest radiographs based on morphological features and wavelet snake model," *Medical Image Analysis*, vol. 6, pp. 431–447, 2002.

[19] H. Kobatake and S. Hashimoto, "Convergence index filter for vector fields," *IEEE Transactions on Image Processing*, vol. 8, no. 8, pp. 1029–1038, 1999.

[20] J. Wei, Y. Hagihara and H. Kobatake, "Detection of cancerous tumours on chest X-ray images: Candidate detection filter and its evaluation," *International Conference on Image Processing 1999*, vol. 3, pp. 397–401, 1999.

[21] C. Pereira, H. Fernandes, A. M. Mendonça and A. Campilho, "Detection of Lung Nodule Candidates in Chest Radiographs," *Lecture Notes in Computer Science*, vol. 4478, pp. 170-177, 2007.

[22] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998.

[23] B. van Ginneken, B. ter Haar Romeny and M. A. Viergever, "Computer-aided diagnosis in chest radiography: A survey," *IEEE Trans. Med. Imag.*, vol. 20, no. 12, pp. 1228–1241, Dec. 2001.

[24] J. Wei, Y. Hagihara, A. Shimizu and H. Kobatake, "Optimal image feature set for detecting lung nodules on chest X-ray images," *Computer Assisted Radiology and Surgery (CARS 2002)*, pp. 706–711, 2002.

[25] C. Pereira, L. Alexandre, A. M. Mendonça and A. Campilho, "A multiclassifier approach for lung nodule classification," *Lecture Notes in Computer Science*, vol. 4142, pp. 612-623, 2006.

[26] G. Coppini, S. Diciotti, M. Falchini, N. Villari and G. Valli, "Neural networks for computer-aided diagnosis: Detection of lung nodules in chest radiograms," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 4, pp. 344–357, Dec. 2003.

[27] K. Suzuki, J. Shiraishi, H. Abe, H. MacMahon and K. Doi, "False positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs by means of massive training artificial neural network," *Acad. Radiol.*, vol. 12, pp. 191–201, 2005.

[28] P. Campadelli, E. Casiraghi and G. Valentini, "Support vector machines for candidate nodules classification," *Neurocomputing*, vol. 68, pp. 281–288, 2005.

[29] P. Campadelli, E. Casiraghi and D. Artioli, "A Fully Automated Method for Lung Nodule Detection From Postero-Anterior Chest Radiographs," *IEEE Transactions on Medical Imaging*, vol. 25, no. 12, pp. 1588 – 1603, Dec. 2006.

[30] R.O. Duda, P.E. Hart and D.G.Stork, *Pattern Classification*, second edition. New York: John Wiley & Sons, 2001.

[31] Z. Jun, L. Xiaomao, L. Jianguo, P. Fuyuan, T. Jinwen, W. Ying, Z. Wenjun and X. Mingxing, "SVM-based ultrasonic medicine image diagnosis," *Proc. SPIE Med. Image Acquisition Process.*, vol. 4549, pp. 92–95, 2001.

[32] S. Xuejun, Q. Wei and S. Dansheng, "Three-class classification in computer-aided diagnosis of breast cancer by support vector machine," *Proc. SPIE Med. Imag.*, vol. 5370, pp. 999–1007, 2004.

[33] S. S. Mohamed and M. A. Salama, “Computer-aided diagnosis for prostate cancer using support vector machine,” *Proc. SPIE Med. Imag.*, vol. 5744, pp. 898–906, 2005.

[34] D. Glotsos, J. Tohka, P. Ravazoula, D. Cavouras and G. Nikiforidis, “Automated diagnosis of brain tumours astrocytomas using probabilistic neural network clustering and support vector machines,” *Int. J. Neural Syst.*, vol. 15, pp. 1–11, 2005.

[35] A. K. Jerebko, J. D. Malley, M. Franaszek and R. M. Summers, “Support vector machines committee classification method for computeraided polyp detection in CT colonography,” *Academic Radiol.*, vol. 12, no. 4, pp. 479–486, 2005.

[36] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, United Kingdom: Cambridge University Press, 2000.

[37] A. Shilton, M. P. Ralph and A. C. Tsoi, “Incremental training of support vector machines,” *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 114–131, 2005.

[38] S. Schölkopf, C. J. Burges and A. J. Smola, *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

[39] J. P. Marques de Sá, *Pattern Recognition: Concepts, Methods and Applications*. New York: Springer, 2001.

[40] R. Akbani, S. Kwek and N. Japkowicz, “Applying Support Vector Machines to Imbalanced Datasets”, *Proc. of ECML 2004*, pp. 39-50, 2004

[41] C. B. Barber, D.P. Dobkin and H.T. Huhdanpaa, "The Quickhull Algorithm for Convex Hulls," *ACM Transactions on Mathematical Software*, vol. 22, no. 4, pp. 469-483, Dec. 1996.