Chapter 3

Segmenting the Lung Fields

Nine algorithms for the automatic delineation of lung fields in standard PA chest radiographs are presented. The algorithms are based on different techniques: matching, pixel classification based on several combinations of features, a rule-based scheme that finds lung contours using a general framework for the detection of oriented edges and ridges in images. Each approach is discussed and the performance of the nine systems on 115 test images is compared with inter-observer variability and results available from the literature. Best performance is obtained by a hybrid scheme that combines the rule-based segmentation algorithm with a pixel classification approach. This combination of two complementary techniques leads to robust performance: the accuracy is above 94% for all images in the test set. The average accuracy of the scheme is $0.969 \pm 0.008$, which is close to the inter-observer variability of $0.984 \pm 0.005$.

CHEST RADIOGRAPHS are of paramount importance in the identification of patients with abnormal pulmonary conditions. A plain chest radiograph frequently also provides a general indication of the type of pathology present. The advent of digital thorax units [74] and digital radiology departments with PACS makes it in principle possible to use computerized methods for the analysis of chest radiographs on a routine basis. In this work, we focus on automated segmentation of lung fields in standard PA chest radiographs. This is a mandatory pre-processing step for most types of computer analysis of chest radiographs. The subject has received a considerable amount of attention in recent literature (see also Chapter 2). The two main approaches are rule-based reasoning and pixel classification with neural networks. With rule-based systems we mean algorithms consisting of a series of steps, each containing specific processing and, usually, certain adjustable parameters. For the segmentation of lung fields, such schemes have been proposed by Xu et al. [269,270], Duryea and Boone [65], and Carrascal et al. [32]. Lung segmentation by pixel classification using neural networks has been investigated by McNitt-Gray [179,178], Hasegawa et al. [103], and Tsujii et al. [235]. Vittitoe et al. [250], [251] developed a pixel classifier for the identification of lung regions using Markov random
field modeling. Brown et al. [28] presented a system to extract lung edges that employs reasoning mechanisms. A third possible approach is to match an input image to segmented reference images. Although matching to atlases has been applied to several problems in medical image processing, the direct application to lung field segmentation has not been explored before.

The fact that statistical classifiers and rule-based schemes seem to be the most popular methods for thorax segmentation is perhaps not surprising because knowledge-based processing is mandatory to solve this task. There are several reasons why the automatic segmentation of chest radiographs is a hard problem from a computer vision point of view. First, there are large anatomical variations in the chest from person to person. Second, the habitus and level of inspiration of the subject during the examination have a profound impact on the image. Third, the settings of the chest unit, particularly the peak tube voltage, determine how well bony structures are visible [10]. But finally, and most importantly, radiographs are projection images and thus contain superimposed structures. Physicians are actually trained to mentally subtract anatomical structures selectively. To cite one popular textbook [225]: “The system generally employed by the radiologist is to look at various structures in a deliberate order, concentrating on the anatomy of each while excluding the superimposed shadows of other structures.” The lung fields in chest radiographs contain several superimposed structures, such as lung vasculature, posterior and anterior ribs, and clavicles. These structures do not make up the borders of the lung fields, as opposed to other structures, such as the mediastinum and the diaphragm. Analysis should differentiate between structures and this can only be done by the incorporation of knowledge.

We consider both a rule-based system and pixel classification with several sets of features. The rule-based scheme we present is new and uses a general and flexible framework for the detection of subdimensional structures, in this case lung edges, about which knowledge is available. Our pixel classifiers use features such as intensity and location, but also features derived from the rule-based analysis and an entropy measure. Furthermore, we propose a new hybrid segmentation scheme that combines the strength of a rule-based approach and pixel classification. We show that this system comes close to the accuracy of segmentation by hand and compares favorably with literature results. Such direct comparisons between different segmentation methods have not been made before.

In the next section we define the exact task and discuss the database we used in our experiments. Subsequently, we will consider each of the three approaches proposed, discuss their advantages and disadvantages and describe the algorithms in detail. In the final sections we present results, discussion and conclusions.

**Purpose, materials and methods**

A possible definition of lung fields is those parts of the image for which the radiation has passed through the lungs. However, this total projected area is impossible to determine from a chest radiograph because of the density of the overlying diaphragm, heart and mediastinum. Therefore, we define the lung fields as those parts of a chest
radiograph which contain lungs not obscured by either diaphragm, mediastinum and heart. It has been estimated \cite{39} that on average 26% of the lung volume and 43% of the total projected area is obscured by one of these structures.

Our database consists of standard PA chest radiographs taken in a tuberculosis screening program for people seeking political asylum in The Netherlands. These images are read by two physicians independently and classified as normal or possibly abnormal. In the latter case, the patient is contacted for further examination. We randomly selected 133 normal and 133 abnormal cases. The images were taken with a mobile Oldelft Electrodelca (Nucletron BV, Veenendaal, The Netherlands), a system commonly used in mass chest screening. The tube voltage was 117 kV and the images were printed on 10 by 10 cm. film. The films were digitized with a Lumisys 100 scanner (Lumisys, Inc., Sunnyvale, CA) to 996 by 996 pixels with 10 bit intensity, subsampled to 256 by 256 pixels, which is sufficient for our purposes. We excluded images from subjects less than 16 years old, and images for which our rule-based segmentation scheme failed to produce any output, indicating gross abnormalities. In these images, 5 in total, at least one of the lung fields was not visible. The remaining 230 images were randomly divided in a test set and a training set, each containing 115 images, about half of which were abnormal. Note that this percentage of abnormal images is much higher than what is encountered in mass chest screening practice. Nevertheless, we did not want to exclude abnormal cases because robust segmentation methods should give good performance on abnormal images as well. Figure 3.1 displays thumbnails of the images to give an idea of the variation encountered in the database.

We indicate the coordinate of the upper left corner of the image with (0,0), and the lower right with (1,1). All references to coordinates and lengths are made in these units, which is convenient because it makes them independent of the image size in pixels. The lung fields were traced with a mouse by two observers (the author (BvG) and his co-worker Bart ter Haar Romeny (BtHR)) independently, under supervision of an experienced radiologist. BvG traced the contours in both the training and the test sets.; these contours are used as the gold standard in the evaluation. BtHR traced all contours in the test set. These contours were used to compute the inter-observer variability.

Approaches to lung segmentation

This section discusses the various possible approaches to lung segmentation.

Matching

Matching, or registration, of images can be defined as finding a transformation that relates points in one image to their corresponding points in another image. Matching can be used for segmentation by matching a segmented reference image to an input image. The transformed reference image yields a transformed segmentation that, if the images are well matched, will be an accurate segmentation of the input image.

A general matching system is determined by four elements \cite{27}. First of all, there is the choice of feature space. This amounts to selecting a reference image, deciding
Figure 3.1: The 115 test images on which segmentation algorithms are evaluated.
on which points in this image to use and possibly applying certain operations on these points, such as filtering. Clearly the choice of a reference image is an important one since the anatomical variation between chest radiographs is large and the selected reference image is the model used during matching. If the reference image is not similar to the input image, up to the transformation searched for, matching will fail. Second, the search space is defined by the parameter space of the allowed transformations. In the case of lung segmentation using reference and input images from different subjects, it is evident that transformations that allow anisotropic scaling are required. One has to choose between global transformations, such as an affine transformation and local, elastic, deformations. Third, a similarity metric is required that should indicate how well the transformed input image and the reference image match. The final element is the search strategy, i.e. the algorithm used to determine the transformation which maximizes the similarity metric. Details about the system used are given in the next section.

An important advantage of this approach is that it incorporates model knowledge and shape information in an implicit way. There is no need for specific analytical models of the objects to be segmented, the reference image is the model.

**Rule-based detection of lung contours**

Rule-based systems offer their designer the freedom to express his knowledge about the problem in any type of rule or processing imaginable. Furthermore, it naturally subdivides the problem in sub-problems. If a part of the system does not yield satisfactory performance, one can add pre- or post-processing steps to correct the problem. There are some obvious weaknesses as well. In practice, rule-based systems consist of a concatenation of steps, each containing several variables, thus leading to an overall scheme that often contains many ad hoc choices and a myriad of user-adjustable parameters. It is usually impractical or impossible to make plausible, let alone prove, that the performance of the system is in some sense optimal. The natural way to improve the system is adding more rules, thus also adding complexity. The same system cannot be applied to a different task: rule-based systems do not generalize.

Another pitfall is the fact that these systems usually take the outcome of one step as input for the next. This may lead to performance that is not robust. In practice, such systems may perform reasonably well for each substep considered separately, but go astray when applied as a whole.

The starting point of our rule-based approach is the observation that the borders between anatomical structures in chest radiographs often largely coincide with edges and ridges in the image. An illustration is given in Figure 3.2. A difficulty is that these structures correspond only partly to the borders between anatomical regions. It is not straightforward to pick out the “correct” structures or parts of structures. Note for instance in the top-right image of Figure 3.2 that the edge between right lung and mediastinum is connected to the edge of the clavicle. We propose a dimensionality reduction technique to overcome some of these problems.

Edges are usually defined as those points in an image where the gradient magnitude is maximum in the gradient direction [31]. Ridges are commonly defined as
Figure 3.2: Edges and ridges in chest images coincide with borders between anatomical structures. A chest radiograph (top-left) with (top-middle) edges at a fine scale of 0.004 and (top-right) a coarser scale of 0.015 and (bottom-middle) ridges (in white) and valleys (in black) at the same fine and (bottom-right) coarse scale. Scale is expressed in image-width units.

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 convoluting the image \( L(x, y) \) with the derivative of a Gaussian \( G(x, y; \sigma) \) at a particular scale \( \sigma \). The normalized Gaussian in 2-D is given by

\[
G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right). \tag{3.1}
\]

If we denote the \( n \)th order derivative in the direction defined by an angle \( \alpha \) as \( L_n^\alpha \) we need to compute

\[
L_n^\alpha(x, y; \sigma) = G_n^\alpha(x, y; \sigma) \otimes L(x, y), \tag{3.2}
\]

where \( \otimes \) denotes convolution and \( G_n^\alpha \) is the \( n \)-th order derivative of the Gaussian kernel in the direction \( \alpha \). Instead of calculating the convolution directly, it is often convenient to express the response for an arbitrary direction \( \alpha \) in terms of combinations of a
finite set of basis filters (such constructions are commonly referred to as \textit{steerable} filters). In particular in 2D, the derivatives of any order \( n \) for any direction \( \alpha \) 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 \( x' \to x \cos \alpha - y \sin \alpha \) and \( y' \to -x \sin \alpha + y \cos \alpha \). Other basis filters are also possible. See [73] for a general discussion on steerable filters. In particular, for order \( n = 1 \) and \( n = 2 \) one obtains

\[
L_1^\alpha(x, y; \sigma) = \cos \alpha L_x + \sin \alpha L_y, \quad (3.3)
\]
\[
L_2^\alpha(x, y; \sigma) = \cos^2 \alpha L_{xx} + \cos \alpha \sin \alpha L_{xy} + \sin^2 \alpha L_{yy}, \quad (3.4)
\]

where on the right-hand side subscripts denote differentiation to \( x \) and \( y \). The detection of extrema is simply performed by comparing each pixel \((x, y)\) with two neighboring locations \((x + \eta \cos \alpha, y + \eta \sin \alpha)\) and \((x - \eta \cos \alpha, y - \eta \sin \alpha)\). If both locations, are lower/higher than the given point, it is a maximum/minimum. We fix \( \eta \) at 0.004 and use linear interpolation.

In this way we obtain structures that cannot cross, because in the direction defined by \( \alpha \) consecutive points cannot both be extreme. This implies that such structures can never “turn”, \textit{i.e.} a line in the direction defined by \( \alpha \) cannot cross a structure more than once. The order of derivative \( n \) and the choice for minima and maxima determines the nature of the detected structures. Using order \( n = 0 \), one detects the axes of bright and dark blobs; \( n = 1 \) yields edges from dark to bright regions, or vice versa; \( n = 2 \) finds bright and dark line structures, all in the direction perpendicular to \( \alpha \). A significance measure for a structure can be computed by integrating \(|L_1^\alpha|\) along the structure or by taking the accumulated length of a structure.

Some results are shown in Figure 3.3. Using the appropriate scales and combinations of directions and derivatives, we extract structures that correspond closely to anatomical structures or boundaries, and which are not connected, such as the structures in Figure 3.2. In Figure 3.3 the thorax center line and the lung center lines are extracted (top-left), the edges of the mediastinum (top-middle), the rib cage (top-right), the rib borders (bottom-middle), and the centerline of costal and intercostal spaces (bottom-right).

The problem of connected structures has greatly been reduced in this way. Since the orientation of lung fields is approximately the same for each radiograph, we can use this knowledge to select the proper directions. Another tunable parameter is the scale \( \sigma \) at which derivatives are computed. In general, scale selection is an active research area, both for single scale and multi-scale methods. In this study we used empirically selected values. One rule of thumb should be that the scale corresponds to the size of the structure to be detected. In practice it may be worthwhile to consider coarser scales since this leaves less structures, which may alleviate the selection of the “correct” structure. To solve this selection problem, we use a straightforward voting technique, the details of which are explained in the algorithm in the next section.

For the detection of diaphragm and lung top, one could use the same techniques. However, the simple voting mechanisms may fail to detect the correct structure. This is caused by, as can be verified from the bottom row of Figure 3.3, the presence of several horizontal edges around the diaphragm and horizontal or circular bright line
structures in the neighborhood of the lung top. Since the lung top and diaphragm are generally the strongest lines and edges, we used dynamic programming [18,3] to detect them. In the algorithm outlined in the next section we will use the fact that the diaphragm corresponds to a maximum in the 1st derivative in the $y$-direction and that the top of the rib cage corresponds to a maximum in the second derivative perpendicular to the rib cage, together with the fact that the contour corresponding to the top of both lungs is approximately circular. Putting this together, estimates of the diaphragm, the border between lung fields and mediastinum, the boundaries of the rib cage and the lung tops are obtained. Figure 3.4 shows representative results.

This algorithm does not necessarily produce the complete contours of both lung fields, since it is not enforced that there are crossings between the mediastinum edge, the lung top, the rib cage, and the diaphragm respectively. In some cases, (parts of) these contours may be missing. We consider this to be a strength of the method, because the fact that structures are not detected may indicate abnormalities. One example is shown in Figure 3.4(a). In fact, this was used to eliminate 5 grossly abnormal images from the database, as mentioned in the previous section. Forcing
Segmenting the Lung Fields

Figure 3.4: (a) Example of an image that is rejected because no thorax center line is found; (b) Typical result of detection of vertical structures (thorax center line, mediastinum edges, lung edges, rib cage); (c) Typical result of detected diaphragm and lung top; (d) Polar transformation of a hemi-circle of the image used in the detection of the lung top. The horizontal axis is the radius, the vertical is the polar angle.

The results of this algorithm are converted to a complete lung contour using a matching approach. We take a lung contour from a reference image and select control points along this contour. These points are moved in horizontal or vertical direction until they coincide with the detected mediastinum, lung top, rib cage or diaphragm respectively. The transformed contours, based on the displacements of control points, define the lung fields. The detected structures are also used to obtain a set of landmark points for each image. These landmarks give an estimate of the width, the height and the location of the center of each lung field and are used to compute features in pixel classification schemes.

Pixel classification

Segmentation can be treated as a pixel classification problem by calculating a feature vector for each pixel in the input image. Output is the anatomical class the pixel belongs to. Any classifier from statistical pattern recognition or neural network theory can be used to approximate this mapping. The classifier is trained with a large set of training samples (pixels from a large collection of training images). Although different types of classifiers will obviously lead to different results, the performance of these segmentation algorithms will depend mostly on the features of the input vector.

One of the main strongholds of this approach is the fact that the results are guaranteed to be optimal in some sense; there are no parameters to be set by a user as compared to the rule-based approach. Consider a simple classifier that only uses intensity as a feature; if the training data is representative of the test data, the classifier will choose the optimal (even multiple) threshold, as opposed to a rule-based thresholding scheme. Another attractive element is the fact that knowledge is implicitly integrated in this approach through the use of training data. Any “general” segmentation technique (like region growing techniques or watershed methods) needs
to be fine-tuned for a particular application, which involves the choice of parameters. In this approach, training the classifiers leads to optimal parameters. Note however, that the actual choice of feature set (and the parameters used to compute them) is, in general, completely arbitrary and a choice of the designer of the system.

In most implementations, pixel classifiers use local features only. It is important to realize the limitations of such systems. It seems impossible, even for human observers, to always correctly classify a very small region as lung or non-lung without the *global* percept of the whole image. Pixel classification approaches have been characterized as region-based and consider rule-based systems as contour-based. This is not necessarily true. One can use contour-based features, such as the distance from a location to a strong edge, in a pixel classification approach. In a rule-based scheme, region measures such as average intensity over a region of interest may be used.

Let us now consider suitable candidate features. The lung fields have a typical shape and location in the image. This suggests pixel location as a feature. The intensity is a good feature since pixels inside the lung fields have lower values than those of the surrounding tissues but higher values than areas outside the body. Furthermore, the lung fields contain overlying ribs and lung vasculature. This yields local variations in intensity at a small scale, while areas outside the lungs are often more homogeneous. We have formalized this notion as follows: subtract a slightly blurred version of the image from the original, thus obtaining a narrow high-pass frequency filtered version of the image; take the absolute value of this image and blur it slightly. This procedure contains two scale parameters and can be considered an entropy measure since it has high response in regions with large local variations. Figure 3.5 shows that this operation indeed produces high output in the lungs, but also at the borders of the film. Finally, the rule-based scheme detects the rib cage, mediastinum edges, diaphragm and lung tops. From the structures detected by the rule-based scheme, a transformation can be computed that scales the width and height of the lungs to standard values (scaling in \(x\)- and \(y\)-direction) and translates the thorax with its center to a standard location. This 4 parameter transformation allows the computation of a *corrected* location for each pixel.
Hybrid scheme

One could argue that pixel classification methods are in some sense complementary to the rule-based method we have discussed previously. This leads to the question whether it is possible to combine both methods. We present a simple method to do this. From all images in the training set, we determine two training sets consisting of all pixels which the rule-based method classified as lung and the pixel classification method as background, and vice versa. We use the corrected location as feature. These sets can be used for reclassification. This hybrid system (System 9) is the most advanced scheme we consider in this paper.

Algorithms for lung segmentation

In this section we will give a detailed description of 9 segmentation schemes that are evaluated in the subsequent experiments.

1. Matching

   We selected one (qualitatively judged as representative) normal thorax as reference. As feature space we used the gray level intensity of the pixels of the image, normalized to zero mean and unit standard deviation and subsampled to 64 by 64 pixels. We allowed affine transformations and used the sum of absolute differences in intensities as similarity metric. We used Powell’s direction set method \cite{202} to find the optimal transformation.

2. Rule-based reasoning

   (a) Detection of thorax center line. The ROI was a rectangular area defined by $0.3 < x < 0.7$ and $0.2 < y < 0.6$. This range for $y$ was also used in steps (b)-(e). The 0th order derivative was computed at a scale of 0.03. Pixels that are maximal in the horizontal direction were detected and grouped into 8-connected structures. For each horizontal line in the ROI, the structure closest to the vertical centerline of the ROI was voted for. The structure that received most votes was selected. Other structures were added, in order of received votes, as long as they did not overlap in their $y$-coordinates. If no center line was detected, the image was rejected.

   (b) Detection of the right/left edge of mediastinum. The ROI was bounded on the right/left by the thorax center line and has a 0.4 width. The 1st derivative in the $x$-direction was computed at a scale of 0.03. Maxima/minima in the horizontal direction were grouped into connected structures. For each horizontal line, starting from the thorax center line, the first structure encountered was voted for. Again, other structures were added, in order of received votes, as long as they don’t overlap in their $y$-coordinates. The same procedure is used in steps (c) to (e).

   (c) Detection of lung center lines. The mediastinum edges are the right and left boundaries of the ROI, which had a 0.3 width. Horizontal minima of the 0th order derivative at a scale of 0.03 were used.
(d) Detection of right/left lung edges. The lung center lines are the right and left boundaries of the ROI which has a 0.4 width. Horizontal minima/maxima of the 1st order derivative in the $x$-direction at a scale of 0.03 were used.

(e) Detection of right/left rib cage. The lung edges are the right/left boundaries of the ROI which has a 0.2 width. Horizontal maxima of the 2nd order derivative at a scale of 0.02 were used.

(f) Detection of diaphragm. The right and left diaphragm were determined with dynamic programming which finds a line from left to right in a rectangular ROI. The ROI contained the 1st derivative in the $y$-direction of the image at a scale of 0.02. The line was allowed to have a slope between -1 and 1. The right diaphragm was detected first, because it is visible more clearly, due to the absence of the heart and stomach bubbles on the left side of the image. The $x$-range of the ROI was from the lowest $x$-coordinate of the rib cage to the lowest $x$-coordinate of the thorax center line. A starting point on the right diaphragm was detected by finding the right lung center at $y = 0.4$ and taking a vertical profile from this point and finding the highest value for the first derivative along this profile. The $y$-range of the diaphragm was 0.4, centered around this point. After the right diaphragm had been detected, crossings with the right rib cage and mediastinum edge were determined. These landmark points were added to the right diaphragm starting point. This resulted in 1, 2, or 3 landmark points. The $y$-range of the ROI for the left diaphragm was given by the smallest $y$-coordinate of these points minus 0.2 of the image height and the largest $y$-coordinate plus 0.2 times the image height. The $x$-range of this ROI was from the highest $x$-coordinate of the thorax center line to the highest $x$-coordinate of the left rib cage.

(g) Detection of lung tops. We used the fact that the top is a bright line structure, just like the rib cage. It is more or less circular, so a polar transformation is made of a hemi-circle in the image to 100 by 100 pixels. As center point the thorax center line at $y = 0.4$ was taken. The radius of the hemi-circle was estimated to be half the distance between the right and left rib cage at this height. The minimum and maximum radius were set at 0.7 and 1.3 times this radius respectively. We took the 2nd derivative along the circle radius and found a maximum path with dynamic programming. In the maximum cost path, we allow deviations from a circular path of at most $45^\circ$.

(h) Lung contours from a reference image with 15 control points along each contour were moved in horizontal or vertical direction until they coincide with the detected mediastinum, lung top, rib cage or diaphragm respectively. The contour interpolated based on the displacements of control points was taken as final lung contour.

(i) Landmark points at half the height and width of the lungs were determined for later use.
3. Pixel classification: most likely class

   This is a trivial classifier in which each pixel was assigned to the most likely class, the background.

4. Pixel classification: location

   In the systems 4 to 9 we used a $k$-nearest-neighbor classifier with $k = 31$. This setting of $k$ was determined as optimal in several pilot experiments. We used 1024 points from each image in the training set (32 by 32). The classification was done for each image in the test set, subsampled to 256 by 256 pixels. In all systems, scaling factors were calculated to obtain zero mean and unit standard deviation for each feature over the whole training set. These scaling factors were also applied to the features of pixels in the test images.

   In system 4 we only used the pixel coordinates as a feature. This yields a classifier that is independent of the input image.

5. Pixel classification: intensity

   In this system we used only pixel intensity as a feature. This results in a system that segments the lung fields by an optimal (multiple) threshold.

6. Pixel classification: intensity and location

   Both intensity and the coordinates of each pixel were used as features.

7. Pixel classification, rule-based reasoning: intensity and corrected location

   Using the landmarks obtained from the rule-based system, we computed a horizontal and vertical scaling factor that scales the estimated width and height of the right lung to fixed values of 0.235 and 0.547. After that, we calculated the translation that takes the center of the thorax to (0.5,0.5). This yielded a corrected location for each pixel that was used as two features ($x$- and $y$-coordinates), together with its intensity.

8. Pixel classification, rule-based reasoning: intensity, entropy and corrected location

   This system was equal to System 6, with an entropy measure added as fourth feature, in addition to intensity and corrected pixel coordinates. The entropy was defined by blurring the image to a fine scale of 0.0039, subtracting this blurred image from the original, taking the absolute value and blurring the result to a scale of 0.0039.

9. Rule-based reasoning and correction using pixel classification

   This system reclassified those pixels in the image for which systems 2 and 7 yielded different classes. There are two possibilities; pixels can be classified as lung by system 2 and as non-lung by system 7, or vice versa. For each case, we determined a training set consisting of all the pixels in the training set for which this situation occurred (subsampled to 256 by 256). We used the corrected location as features.
Chapter 3

Results

Segmentation performance can be measured in various ways, and ultimately, the requirements of the application that uses the segmentation as input determine whether the segmentation is sufficiently accurate. Here we consider the problem as segmentation between lung and background and calculate the classical accuracy, sensitivity and specificity

\[
\text{Accuracy} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}},
\]

(3.5)

\[
\text{Sensitivity} = \frac{N_{TP}}{N_{TP} + N_{FN}},
\]

(3.6)

\[
\text{Specificity} = \frac{N_{TN}}{N_{TN} + N_{FP}},
\]

(3.7)

to measure segmentation performance. \(N_{TP}\) is the true positive fraction (part of the image correctly classified as lung), \(N_{TN}\) is the true negative fraction (part of the image correctly classified as background), \(N_{FP}\) is the false positive fraction (part of the image incorrectly classified as lung), and \(N_{FN}\) is the false negative fraction (part of the image incorrectly classified as background).

The contours manually traced by BvG were used as gold standard. The lung contours in the test set were manually segmented by BtHR, and from these contours the inter-observer variability was computed.

In the literature, other methods are used to evaluate the performance of segmentation schemes. One popular approach is to present the results to radiologists who rate the performance on a qualitative scale. A drawback of this method is that it is impossible to compare results of such evaluations with other studies. Accuracy, sensitivity and specificity have been used, or can be calculated from results reported, in several literature studies.

Table 3.1 lists the results of our schemes together with results reported in the literature. The results are sorted according to accuracy. Key result is the accuracy of 96.9% of system 9, which approaches the inter-observer variability of 98.4%. Another interesting result is the fact that we did not find significant differences in segmentation accuracy for normal versus abnormal radiographs.

Figure 3.6 shows results of several systems for 5 images. The images were chosen to range from best to worst performance for System 9, the overall most accurate system. Figure 3.7 shows the fraction of false positives and false negatives as a function of location for systems 8, 2 and 9.

All experiments were performed on a standard PC with a 450 MHz Pentium III processor. The rule-based segmentation (system 2) required 4.1 seconds on average. Pixel classification schemes required anywhere from 8 seconds to 4 seconds. The reclassification in system 9 requires the computation time of both the rule-based scheme, the pixel classifier of system 7, and reclassification. In total, this took 10.4 seconds on average. Segmentation by matching took about 24 seconds.
### Table 3.1: Accuracy, sensitivity and specificity of all 9 systems on the complete test set. PC stands for pixel classification, MRF for Markov random field model. The number in parentheses denotes the system number. The systems are sorted according to accuracy. Below the horizontal line are results from those literature studies for which the same performance measures were reported (or could be computed). Note that the test sets in these studies are different.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-observer variability</td>
<td>0.984 ± 0.00475</td>
<td>0.957 ± 0.0174</td>
<td>0.993 ± 0.00306</td>
</tr>
<tr>
<td>Classification correction (9)</td>
<td>0.969 ± 0.00803</td>
<td>0.943 ± 0.0330</td>
<td>0.978 ± 0.0106</td>
</tr>
<tr>
<td>Rule-based (2)</td>
<td>0.961 ± 0.0116</td>
<td>0.940 ± 0.0389</td>
<td>0.969 ± 0.0153</td>
</tr>
<tr>
<td>PC int., entropy, corr. location (8)</td>
<td>0.956 ± 0.0157</td>
<td>0.912 ± 0.0617</td>
<td>0.972 ± 0.0248</td>
</tr>
<tr>
<td>PC int., corrected location (7)</td>
<td>0.953 ± 0.0177</td>
<td>0.906 ± 0.0649</td>
<td>0.970 ± 0.0288</td>
</tr>
<tr>
<td>PC int. and location (6)</td>
<td>0.933 ± 0.0219</td>
<td>0.854 ± 0.0771</td>
<td>0.966 ± 0.0309</td>
</tr>
<tr>
<td>Matching (1)</td>
<td>0.907 ± 0.0437</td>
<td>0.766 ± 0.153</td>
<td>0.946 ± 0.0480</td>
</tr>
<tr>
<td>PC location (4)</td>
<td>0.898 ± 0.0382</td>
<td>0.784 ± 0.0984</td>
<td>0.947 ± 0.0339</td>
</tr>
<tr>
<td>PC int. (5)</td>
<td>0.847 ± 0.0356</td>
<td>0.727 ± 0.122</td>
<td>0.891 ± 0.0447</td>
</tr>
<tr>
<td>All negative (3)</td>
<td>0.736 ± 0.0551</td>
<td>0 ± 0</td>
<td>1 ± 0</td>
</tr>
<tr>
<td>Duryea: rule-based method [65]</td>
<td>0.959 ± 0.054</td>
<td>0.863 ± 0.11</td>
<td>0.987 ± 0.044</td>
</tr>
<tr>
<td>Vittitoe: MRF [250]</td>
<td>0.948 ± 0.016</td>
<td>0.907 ± 0.044</td>
<td>0.972 ± 0.020</td>
</tr>
<tr>
<td>McNitt-Gray: 59 features [178]</td>
<td>0.932</td>
<td>0.949</td>
<td>0.922</td>
</tr>
<tr>
<td>Tsujii: PC corr. int., loc. [235]</td>
<td>0.923</td>
<td>0.903</td>
<td>0.930</td>
</tr>
<tr>
<td>McNitt-Gray: 8 features [178]</td>
<td>0.918</td>
<td>0.903</td>
<td>0.930</td>
</tr>
<tr>
<td>Vittitoe: PC int. 3x3 [250]</td>
<td>0.893 ± 0.027</td>
<td>0.846 ± 0.057</td>
<td>0.925 ± 0.038</td>
</tr>
<tr>
<td>Vittitoe: PC location [250]</td>
<td>0.880 ± 0.035</td>
<td>0.820 ± 0.08</td>
<td>0.920 ± 0.044</td>
</tr>
<tr>
<td>Duryea: classifying location [65]</td>
<td>0.879 ± 0.063</td>
<td>0.785 ± 0.10</td>
<td>0.934 ± 0.063</td>
</tr>
<tr>
<td>Vittitoe: fixed thresholding [250]</td>
<td>0.806 ± 0.071</td>
<td>0.860 ± 0.058</td>
<td>0.781 ± 0.112</td>
</tr>
<tr>
<td>Duryea: all negative [65]</td>
<td>0.751 ± 0.063</td>
<td>0 ± 0</td>
<td>1 ± 0</td>
</tr>
</tbody>
</table>

**Discussion**

We start with a discussion about the performance of the “simple” systems. Assigning all pixels to the most likely class, the background, yields a 73.6% accuracy. Duryea, who calculated the same number on his dataset of 802 images, reported a background percentage of 75.1. Taking the most likely class as a function of location (segmentation independent of the input image) gives an accuracy of 89.8%. The same method was applied by Duryea and Boone and by Vittitoe et al. and they reported 87.9% and 88.0%, respectively. The similarity of these results, although different image sets were used, is striking.

Our result using only intensity as feature, which can be seen as using an optimal multiple threshold, yields an accuracy 84.7%. Vittitoe determined a single optimal threshold for his dataset, which scored 80.6%. Using a neural network classifier with the intensities of a 3 by 3 neighborhood (with the images subsampled to 64 by 64 pixels) of each pixel as input vector, he achieved 89.3% correct classification. The fact that our result with intensity as a single feature scores half-way between those results indicates that the performance increase of the 3 by 3 intensity classification versus fixed thresholding is partly due to the fact that a classifier determines multiple
thresholds and partly due to the textural context information obtained from the pixel neighborhood.

The result of the matching system, accuracy 90.7%, is disappointing, since it hardly outperforms the “average lung field” method. In principle, the matching system can be improved and expanded in several ways. Clearly, a different reference image would give different results. One could test a range of reference images and use the one with the best overall performance. If some information about the chest radiograph is known beforehand, such as the age and gender of the subject, one could use a reference image with similar characteristics. Since the reference image should contain features common to all chest radiographs but not those particular to just one, one could take the average image of a range of matched chest images. A more involved alternative would be to do the matching with several reference images and keep the one with the highest similarity measure as the final result. However, pilot experiments with that approach did not lead to much improvement. Varying matching parameters such as the mask size (feature space) and using different similarity

![Segmentation results](image)

**Figure 3.6:** Segmentation results for 5 radiographs, shown in the left column, for several systems, indicated in parentheses under each column. The right column shows the inter-observer variability (a comparison of the segmentations of two observers). True negative pixels are shown white, true positive light gray, false positive pixels are dark gray and false negative pixels are shown in black. All 115 test images were sorted based on the accuracy of system 9, the most accurate system and assigned a position, 1 for the highest accuracy, 115 for the lowest accuracy. The top row shows image 1 (highest accuracy), the bottom row shows image 115 (lowest accuracy). The second to fourth row show the images at positions 23, 57, and 80, respectively.
metrics, such as mutual information and histogram energy, gave similar results in all cases. In principle definite performance improvement could be obtained by allowing elastic transformations. It can be seen easily by inspecting some chest radiographs that even an optimal affine transformation will not result in a perfect match between images.

Now let’s consider the inter-observer variability. This score provides a theoretical upper bound. The inter-observer variability is mainly due to the difficulty in assessing the exact borders of mediastinum. Given the performance of “simple” systems and the upper bound set by the inter-observer variability, a segmentation scheme for lung fields in chest radiographs will in practice have an accuracy between 90% and 98%.

It is surprising that a classifier using intensity and the location of pixels as feature (system 6) already gives an accuracy of 93.3%, especially if this is compared with the results of McNitt-Gray [178] who obtained 93.2% using 59 features and spatial information. We do not think that our use of a $k$NN-classifier instead of a neural network causes the good performance of our pixel classifiers; McNitt-Gray concluded [179] that neural networks only slightly outperformed $k$NN-classifiers on his data. It may be partly explained by the difference in database. McNitt-Gray used a test and training set of only 17 images each and our much larger database may contain more statistical information. However, it may also indicate that, apart from spatial information, the use of many more local features than (raw) intensity, can hardly improve the discriminating power of pixel classifiers in segmenting lung fields.

There are several arguments in favor of this hypothesis. Our rule-based scheme, and Duryea’s rule-based scheme, outperforms McNitt-Gray’s 59 feature classifier and Vittitoe’s pixel classification using Markov random field modeling. In this study, little improvement is obtained by adding entropy as a feature to our classifier; the accuracy of system 8 is only 0.3% higher than that of system 7. And when using 8 features out of his total set of 59, the performance of McNitt-Gray’s classifier did not decrease much.

The rule-based system (system 2) is accurate and robust, but there is still room for improvements. As to be expected, adding rules seems the straightforward way to higher accuracy. We distinguish two types of error: occasional and structural.
The false positive and false negative maps, Figure 3.7c and 3.7d, provide a good way to locate structural errors. It is evident that the detected edge of left lung and mediastinum in the upper lung part is structurally too much to the right. Note that the false positive and false negative maps for systems based on classifiers 3.7(a-b) and 3.7(e-f) do not show such pronounced locations of structural errors. This is an example of a situation where classifiers benefit from their guaranteed optimal performance, contrary to rule-based schemes where the intuition of the designer may fail. Other structural errors are the detection of the outer rib cage, which is detected as a line structure and located on the center of the overlapping ribs. The lung fields end slightly more medial. Because of the coarse scales used to detect the lung edges, their location has shifted. In principle, one could use edge focusing techniques [19] to trace the edges back to a finer scale. In practice, this procedure introduces new scale parameters and does not easily lead to improved results, especially not if the edges are curved or connected.

The most important categories of occasional errors, of which examples are given in Figure 3.8, are

- Complicated patterns of stomach gasses and cardiac edges are common near the lower left lung fields. Therefore the detection of the left diaphragm occasionally fails. This occurred in 18 of the 115 cases. The failure may have profound effects, as shown in Figure 3.8(top-left), or minor consequences (Figure 3.8(top-right)).

- Another common situation is dense lower lung fields, usually due to poor inspiration of the subject. In 9 cases this leads to failure to detect the left diaphragm (Figure 3.8(bottom-left)).

- In two cases, the lung tops were not accurately detected. This was caused by poor visibility of the rib cage. An example is shown in Figure 3.8(bottom-middle).

- Large hila or dense medial lung fields may cause inaccuracies in the detection of the mediastinum edges. Partly this is a structural problem as well. An example is shown in Figure 3.8(bottom-right).

Increasing the robustness of the left diaphragm detection could significantly improve the segmentation method. Indications for failure are differences between the average height of right and left diaphragm and the presence of diaphragm-like structures above/below the detected diaphragm (in case of unexpectedly low/high left diaphragms). Adding rules is the obvious way to extend a rule-based system and this has been noted by Xu [270] who developed tests to classify thoraxes into several categories based on the stomach gasses appearance. Unfortunately, he did not evaluate his scheme in terms of accuracy, sensitivity and specificity, so we cannot compare the performance of his algorithms with ours. Rule-based reasoning is a proper tool to tackle this problem. Vittitoe [250] states that textural features may discriminate between gassy stomach and lung regions. Although this might be true in some cases, in general we expect that local analysis will be unable to make the distinction because through the stomach gasses, one discerns lung texture.
Figure 3.8: Overview of occasional failures of the rule-based segmentation scheme. (Top row) Largely and slightly incorrect detection of left diaphragm due to the presence of stomach gasses (occurred in 18 of 115 cases). (Bottom-left) Incorrect detection of left diaphragm due to a dense low left lung field (2 cases). (Bottom-middle) Incorrect detection of the lung top (9 cases). (Bottom-right) Incorrect detection of the right mediastinum edge due to a pronounced right hilum.

The remaining pixel classifiers (Systems 7 and 8) use the corrected location as feature. Since these locations are calculated based on the results of the rule-based scheme, these systems combine rule-based results with pixel classification. Nevertheless, their performance does not exceed that of the rule-based method. System 8 is comparable to the approach of Tsujii [235], who used a corrected location, an intensity feature and an entropy measure. The fact that he obtains lower accuracy may be due to a different database, but we must also note that his scheme to determine the corrected location is rather simple, compared to our approach. The brightness of the radiographs varies per subject (although the Electrodelca chest unit performs some correction to obtain equally bright images), one could attempt to correct for this by scaling the intensity values relative to certain areas that are assumed to be in the lung fields or in the mediastinum. This was done by Tsujii. Another alternative could be to use the entry of a pixel in the cumulative histogram, instead of its intensity as feature. McNitt-Gray investigated the use of this feature. This is equivalent to replacing radiographs with histogram equalized versions. To us this seems rather odd since the amount of dark areas outside the body, varies significantly per image, and affects where pixels within the lung fields end up in the cumulative histogram.
In pilot experiments, we did not note any performance gain using such corrections.

Using corrected location alone is not sufficient to outperform the rule-based method with a classifier. Therefore we attempted to combine the strengths of both methods in a system that reclassified those pixels for which rule-based segmentation and pixel classification with corrected location and intensity gave different results. This (system 9) increased the accuracy from 0.961 to 0.969. This improvement may seem small, but in fact it bridges 35% of the gap between the rule-based scheme (96.1%) and the maximally achievable inter-observer variability of 98.4%, and therefore may be considered a significant improvement. The result can be appreciated in Figure 3.7. Note that the structural errors of the rule-based system (Figure 3.7(d)) have decreased while at the same time the overall performance is better than that of the pixel classifier of system 8 (Figure 3.7(a) versus 3.7(e) and 3.7(b) versus 3.7(f)).

Even more important than the increase in average accuracy, sensitivity and specificity, is the gain in robustness. The standard deviation drops by 30% relative to the rule-based system. Reclassification is a means to correct those cases where the rule-based system makes an occasional gross failure. If we consider the 10% of images that scored worst with the rule-based scheme, the average accuracy is 93.7%. For the corrected scheme this percentage is 95.2%. Note that these are different images, so for the 10 worst cases for the rule-based method, the performance gain is even larger. The worst case of all 115 images using the rule-based system scored 90.1%, for the corrected scheme the worst result was 94.0%. The increase in robustness is also visualized in Figure 3.9, where one can observe that the difference in accuracy between the corrected scheme and schemes 2 and 7 is largest for those images that are relatively difficult to segment.

Note that only pixels classified differently by systems 2 and 7 are eligible for reclassification. Clearly this does not include all incorrectly classified pixels, and thus this system may be improved by using a more elaborate reclassification scheme. Another improvement could be to take into account the fact that lung regions cannot contain holes or ragged edges, which can result from pixel classification. This could be implemented using a final matching step on the result of system 9, just as was done in order to convert the contours detected by the rule-based system into closed lung field contours.

The implementations we used were not aggressively optimized for speed, although the kNN-implementation by Mount and Arya [8] available at http://www.cs.umd.edu/~mount/ANN is very efficient. There is ample room for speed improvements. First of all, subsampled versions of the input images can be used, after which for higher resolution images only pixels close to the lung contours need to be classified. For the rule-based scheme, 90% of the computation time is needed to compute the various derivatives of the input image. For most derivatives, these computations are only required for a small part of the image. An implementation that takes advantage of this fact, and possibly uses fast (approximate) implementations of derivatives, will be several orders of magnitude faster. Furthermore, the classification times can be reduced by allowing approximate nearest neighbors [8] or even using look-up tables. The hybrid scheme uses both the rule-based segmentation and the pixel classifier from system 7. These two calculations are independent and could be implemented in parallel on a dual processor PC. Taken this together, the segmentation of a single
input image, even at a full resolution used in clinical practice, can be done within a second. This allows for application in clinical and mass chest screening practice.

**Conclusions**

One of the goals of this study was to compare the performance of various approaches to the segmentation of lung fields in chest radiographs. Because of the detailed comparisons between the 9 systems considered here, we believe we may have shed light on which parts and elements of the segmentation schemes proposed in this work and previous studies are actually important for their performance in practice. It turns out that both pixel classification with simple features and rule-based schemes focused on contour detection can be used to obtain good results. All in all, rule-based lung contour detection seems to perform slightly better than pixel classification. This indicates that the use of global information, to which pixel classification approaches are not directly suited, is important for lung field segmentation. The rule-based scheme we have developed in this work is robust and accurate, but can still be improved in several ways, notably by adding rules to handle exceptional situations. We have shown that both methods are complementary, because they can be combined into a hybrid scheme that is both more robust and more accurate. This accuracy of this
scheme approaches the theoretical limit of inter-observer variability and seems to outperform algorithms published so far, although this remains to be proven by an extensive evaluation of several algorithms on one, common database.