Geometrical and texture features estimation of lung cancer and TB images using chest X-ray database

S.A. Patil*

Textile and Engineering Institute,
Ichalkaranji 416115, Maharashtra, India
E-mail: shrinivasapatil@gmail.com
*Corresponding author

V.R. Udupi

Maratha Mandal's College of Engineering,
Belgaum, 59001, Karnataka, India
E-mail: vishwa_u@yahoo.com

Abstract: Early detection is the most promising way to enhance a patient's chance for survival of lung cancer. Detection of disease through image is one of the most challenging tasks in medical image analysis. A computer algorithm for nodule detection in chest radiographs is presented. The algorithm includes four main steps like, image acquisition, image pre-processing, nodule detection, and feature extraction. Total 75 lung cancer and TB images are used during experiment to estimate geometrical and texture features. Manual lung field segmentation with Gray Level Co-occurrence Matrix (GLCM) techniques are used for segmentation and texture features estimation respectively.

Keywords: chest X-ray; active shape modelling; GLCM; gray level co-occurrence matrix; lung field segmentation.

Reference to this paper should be made as follows: Patil S.A. and Udupi, V.R. (2011) 'Geometrical and texture features estimation of lung cancer and TB images using chest X-ray database', *Int. J. Biomedical Engineering and Technology*, Vol. 6, No. 1, pp.58–75.

Biographical notes: S.A. Patil obtained his BE in Electronics Engineering from Shivaji University, Kolhapur, India, in 1988 and MTech in Bio-medical Engineering from IIT Bombay during 1997. Presently, he is working as an Assistant Professor in Electronics Engineering Department of Textile and Engineering Institute, Ichalkaranji, Maharashtra. He has registered for PhD in Shivaji University and his research topic is 'A novel neural network approach to detect and classify lung cancer using chest radiographs'. He has a number of publications in journals and conferences.

V.R. Udupi has 25 years of teaching experience and presently working as a Principal at Maratha Mandal's College of Engineering, Belgaum, Karnataka, India. He has completed his Graduation in Electronics and Communication Engineering from Mysore University in 1984 and Post-Graduation in Electronics Engineering from Shivaji University, Kolhapur, India, in 1989.

He has obtained PhD from Shivaji University in Electronics Engineering during 2003. He has over 42 papers, 2 books, 8 PhDs to his credit. He is a life member of ISOI, CSI, SSI and BMESI.

1 Introduction

Lung cancer is one of the most common and deadly diseases in the world. The prognosis and the cure of lung cancer depend highly on the early detection and treatment of small and localised tumours. As reported by Heelan, Brett and Nash, the detection of lung tumours in the early stage of growth can result in a better prognosis for survival. The five-year patient survival rate is approximately 40% when lung cancer is detected in the early stage. About 87% of lung cancers are thought to result from smoking or passive exposure to tobacco smoking.

Physical characteristics of the nodules, such as rate of growth, pattern of calcification and type of margins, are very important in the investigation of the solitary lung nodules. Every lung nodule grows in volume over time. However, malignant nodules grow at an exponential rate, which is usually expressed as a tumour's doubling time. Malignant nodules have a doubling time of between 25 and 450 days, whereas the benign nodules are stable and have a doubling time more than 500 days. In addition to the rate of growth of the nodules, the pattern of the calcification is an important indicator of whether the nodule is benign or malignant. Nodules that are centrally or diffuse calcified are usually benign. Manual lung nodule detection, which was possible using chest X-ray, is no longer possible. It is necessary to have automated tools that can assist a physician in quickly detecting the nodules.

Chest X-ray image has been used for detecting lung cancer for a long time. The early detection and diagnosis of pulmonary nodules in chest X-ray image are among the most challenging clinical tasks performed by radiologists. Some of these lesions may not be detected because they may be camouflaged by the underlying anatomical structures, or the low quality of the images, or the subjective and variable decision criteria used by the radiologist.

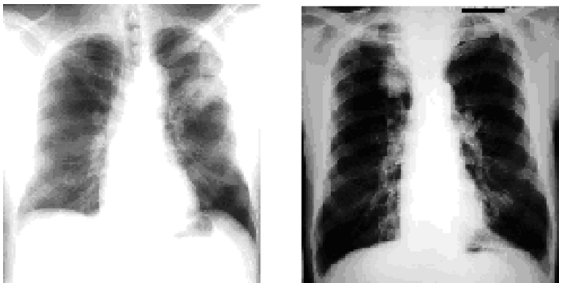
Computer-Aided Diagnosis (CAD) has been proven to be a very effective approach as assistant to radiologists for improving diagnostic accuracy. Numerous systems were reported for detecting lung nodules on chest X-ray images. However, the strong concern of almost all of them is that the false positives per image are too large. How to reduce the number of false positives while maintaining a high true positive detection rate is the most important work in realising a chest CAD system. Most of the proposed CAD systems adopt a two-step pattern recognition approach, which is a combination of a feature extraction process and a classification process using neural network classifier or statistical classifier. The performance of the classifier depends directly on the ability of characterisation of candidate regions by the adopted features. Many kinds of features have been proposed for discriminating between normal tissues and abnormal ones. However, there have been a few researches on comparing the effectiveness of those features. The purpose of our research is to find the optimal feature set from the sufficiently available database for the classification.

2 Features

Most of the lung cancers start in the lining of the bronchi. Less often, cancers begin in the trachea, bronchioles, or alveoli. Lung cancers are thought to develop over a period of many years. As a cancer develops, the cancer cells may produce chemicals that cause new blood vessels to form nearby. These new blood vessels nourish the cancer cells, which can continue to grow and form a tumour large enough to see on X-rays. Cells from the cancer can break away from the original tumour and spread to other parts of the body. As noted earlier, this process is called metastasis.

Mainly, the lung cancers are classified as Small-Cell type of Lung Cancer (SCLC), and Non-Small-Cell type of lung cancer (NSCLC) (Figure 1). Usually, SCLC arises at alveolar level or at terminal bronchial level, and seen to be more scattered in nature on X-ray. NSCLC arises in the larger, more central bronchi; tends to spread locally; metastasises somewhat larger than the other patterns, but its rate of growth in its site of origin is usually more rapid than that of other types.

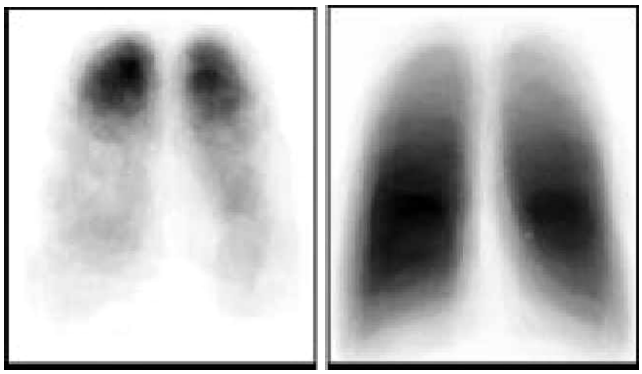
Figure 1 SCLC and NSCLC images



Tuberculosis (TB), though no longer the ubiquitous and sinister threat of past decades, is still a problem that must be borne in mind, particularly with immigrants from developing countries and in the immune-suppressed patients. The primary form of TB infection, which is used to be seen almost exclusively in children, is now also being seen in older patients. The primary TB infection appears on X-ray as a diffuse opacity representing a patch of consolidation in the lung field with increased striations extending towards the hilum where the enlarged glands show as rounded opacities. Pleural effusion is also a common manifestation of primary TB infection. The secondary or adult type of TB infection mostly affects the posterior segment of the upper lobe. On the Posterior–Anterior (PA) film, it appears as an area of shadowing near the lung apex often mottled in character. Figure 2 gives an idea of the spatial distribution of the abnormal areas within the chest radiographs in the TB database.

Feature selection is a very important step in organising a classifier. Theoretical approach cannot be applied to determine the optimal combination of features, and the only way to select the optimal feature subset is to evaluate all possible combinations of the features. Moreover, sufficient numbers of test materials are necessary to evaluate the performance of each feature combination. It means that the number of combinations and the total amount of computation time become impractically huge. Therefore, Jun Wei, Yoshihiro Hagihara, Akinobu Shimizu, and Hidefumi Kobatake accepted heuristic algorithms such as a genetic algorithm, a forward stepwise and a backward stepwise selection technique to decide the optimal feature set.

Figure 2 (a) Distribution of abnormal areas for the TB database and (b) distribution of abnormal areas for the Interstitial Disease (ID) database



(a) (b)

According to Wei et al. the features are classified as

2.1 Geometric features

Spreadness, Circularity, Area, Equivalent radius, Distance from the candidate point to the pulmonary hilum and Flatness are some of the geometric features. Such geometric features are calculated from the binary Suspicious Region (SR) using thresholding technique.

2.2 Texture or contrast features

Generally, tumour region is brighter than its background on X-ray image. So, the contrast information can be used as features. Contrast features are again classified under two categories, first-order statistic and second-order statistic. In this work, such 10 kinds of features are calculated from SRs.

2.2.1 First-order statistics features

First-order statistics are calculated from histograms of the grey-scale values. The histograms are obtained from filtered images. Features calculated from each histogram include average grey level, standard deviation, contrast, skewness, kurtosis and entropy.

2.2.2 Second-order statistics features

Co-occurrence matrix method has been adopted to extract features of second-order statistics. They are obtained by using Haralick transformation. Co-occurrence matrices are obtained from the inner and the outer regions of each SR. Correlation, energy, homogeneity and contrast are the features computed by using co-occurrence matrix.

3 Related work

Toriwaki et al. (1973) have carried out the first work in this field to detect Suspected Nodule Areas (SNA) with the help of image processing technology. But the system does not give very accurate results, as they only detect nodules from approximately 1 cm. In recent research, it has been found that, the difference-image technique originally proposed by Giger et al. (1988, 1990), or a variant thereof (e.g., Carreira et al., 1998; Keserci and Yoshida, 2002), is used. In the difference-image technique, the original image is filtered twice: Once with spherical kernel to obtain a nodule-enhanced image, and once a median filter is applied to obtain a nodule-suppressed image. Nodule candidates are then obtained by thresholding the subtraction of the two filtered images.

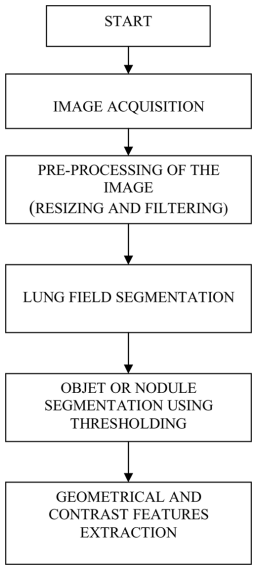
However, the largest part of the variability between published models resides in the feature extraction step. Here, the methods use different sets of meaningful characteristics that ought to enable the classifier to distinguish between actual nodules and False Positives (FPs). These FPs include rib crossings, rib–vessel crossings, vessel–vessel crossings and end-on vessels. During 1990–1996, several methods have been proposed to reduce the number of FPs while maintaining high sensitivity. Most of them operate in two phases, feature extraction and feature classification. Some morphology-based algorithms have been proposed to extract specific features, such as circularity, size, contrast, or local curvature. Used feature sets are mostly derived from histogram information (Sklansky and Petkovic, 1984; Giger et al., 1990), filter outputs (e.g., Keserci and Yoshida, 2002), or descriptions of candidate shape (Sankar and Sklansky, 1982; Carreira et al., 1998; Li et al., 2001). Wei et al. (2002) investigated the feasibility of finding an optimal set of features, derived as a subset from sets of three classes of features. Yue and Gosntasby (1995) have developed an algorithm for detection of posterior rib borders, which is one of the FPs, in chest radiographs. Zhang and Huang (1997) have developed Automatic Background Recognition and Removal (ABRR) method for chest radiographs. This is one of the features considered while doing segmentation with the help of image processing. ABRR has shown an excellent performance (99%) of correct recognition and removal of (91%) the background signal. Penedo et al. (1998) have developed CAD system, which is capable of detecting big nodules (not better for small nodules) when they are in initial stage. Van Ginneken et al. (2002) presented a fully automatic scheme for texture analysis of lung fields in chest radiographs. The method is based on texture analysis on local region in the image.

4 Method

The steps followed for analysis and feature extraction from lung cancer X-ray images are presented in Figure 3. As an initial step, the images are obtained using image acquisition method and then the application of pre-processing algorithms, including size normalisation and filtering of the image. The features, those are identified to be useful for diagnosis and analysis, require separation of the lung fields from the background. Lung field masks are prepared manually by segmenting the lung fields, as well as readily available masks, developed by using ASM technique, are used to separate the lung fields (refer JSRT public database). Thresholding along with region-based segmentation techniques are used to segment the lung nodules (in case of NSCLC images) and cancerous portion (in case of SCLC images) from the separated lung field area.

And in the next step, geometrical and texture features, discussed in the above-mentioned section, have been estimated. The next section describes the detail description of the selected features as well as the methodology used to extract them from the PA chest X-ray images.

Figure 3 Flow of the feature extraction technique



The chest unit used for screening X-ray films is mobile KlinoskopH unit (Siemens, India make). Keeping the tube voltage equal to 150 kV, 500 mA at 2.2 mm Pb the images were printed on 14×17 cm^2 film. Then, these films were digitised with a high-resolution scanner (Scanjet 2400, HP India make). These films were collected from the private medical institutes. Lung cancer and TB images from the public database are also used in this study. Every image data is acquired with 256 grey levels (8 bits) and stored as JPEG (.jpg, .jpeg) data. Before extraction of the features from an image, it is necessary to pre-process the image to reduce irrelevant information or noise, and to enhance the image properties, which makes the feature measurement easier and more reliable. Scanned images are resized to a size of 512×512 pixels. Median filter is used to remove the noise or irrelevant information from the images.

The purpose of the segmentation is to find corresponding regions within the lung fields. Segmentation of lung fields on PA chest radiographs has received considerable attention in the literature.

Rule-based schemes have been proposed by Armato et al. (1998), Xu and Doi (1995), Duryea and Boone (1995), Pietka (1994) and Brown et al. (1998). Lung segmentation by pixel classification using neural networks has been investigated by McNitt-Gray et al. (1995) and Tsujii et al. (1998). Vittitoe et al. (1998) developed a pixel classifier for the identification of lung regions using Markov random field modelling. Van Ginneken and ter Haar Romeny (1995) proposed a hybrid method (improved Active Shape Modelling Technique) that combines a rule-based scheme with a pixel classifier.

Active Shape Model, a general technique for image segmentation, has been developed by Cootes and Taylor (1999) and has been applied to various segmentation

tasks in medical imaging. Below, the ASM scheme is described briefly. An object is described by points, referred to as landmark points. The landmark points are (manually) determined in a set of training images. From these collections of landmark points, a point distribution model is constructed as follows. The landmark points (x_1, y_1), …, (x_n, y_n) are stacked in shape vectors.

$$X = (x_1, y_1, \ldots, x_n, y_n)^T. \tag{1}$$

Given a set of shapes, Principal Component Analysis is applied to these vectors (without performing any kind of alignment), by computing the mean shape and the covariance

$$X' = \frac{1}{s}\sum_{i=1}^{n} X_i \tag{2}$$

$$S = \frac{1}{s-1}\sum_{i=1}^{n} (X_i - X')(X_i - X')^T \tag{3}$$

and the eigensystem of the covariance matrix. The eigenvectors corresponding to the t largest eigenvalues λ_i are retained in a matrix, $\Phi = (\Phi_1 \mid \Phi_2 \mid \Phi_3 \mid \ldots \mid \Phi_t)$. A shape can now be approximated by,

$$X \approx X' + \Phi b, \tag{4}$$

where b is the vector of t elements containing the model parameters, computed by

$$b = \Phi^T (X - X'). \tag{5}$$

When fitting the model to a set of points, the values of b are constrained to lie in the range of several times $\pm\sqrt{\lambda_i}$. The number t of eigenvalues to retain is chosen so as to explain a certain proportion f_v of the variance in the training shapes. The desired number of modes is given by the smallest t for which

$$\sum_{i=1}^{t} \lambda_i \geq f_v \sum_{i=1}^{2n} \lambda_i. \tag{6}$$

To create models of the image profiles around each landmark, profiles g_1, …, g_n are sampled around each landmark, perpendicular to the contour. Sampling k pixels on either side of the profiles gives profiles of length $2k + 1$. The first derivatives of these profiles are used. The profiles are also normalised by dividing through the sum of absolute values of the elements. For each landmark point, the mean profile g' and the covariance matrix S_g are computed. To fit the model, the Mahalanobis distance between a new profile g_i and the profile model can be computed.

$$f(g_i) = (g_i - g')\, S_g^{-1} (g_i - g'). \tag{7}$$

Minimising this Mahalanobis distance $f(g_i)$ is equivalent to maximising the probability that g_i originates from the distribution $\{g_1, \ldots, g_n\}$.

This minimisation is used to find new locations for the landmarks during fitting.

These profile models, given b, g' and S_g, are constructed for multiple resolutions. The finest resolution uses the original image and a step size of 1 pixel when sampling

the profiles. The next resolution is the image observed at scale $\sigma = 1$ and step size of 2 pixels. Subsequent levels are constructed by doubling the image scale and the step size.

Shapes are fitted in an iterative manner, starting from the mean shape. Each landmark is moved along the direction to the contour to n_s positions on either side, evaluating a total of $2n_s + 1$ position. The landmark is put at the position with the lowest Mahalanobis distance. After moving all landmarks, the shape model is fitted to the points (Figure 4), yielding an updated segmentation. This is repeated a fixed number of N times at each resolution, from coarse to fine. Around 180 lung field masks (see Figure 5) prepared using ASM technique are available (refer JSRT and SCR public database) for lung field segmentation (separately for each right and left lung fields).

Figure 4 Landmark points with lung fields

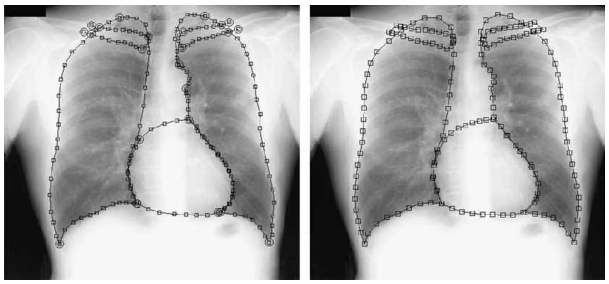
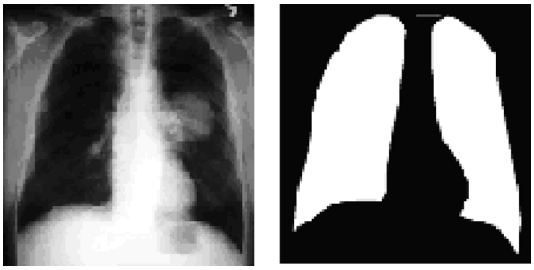


Figure 5 Original X-ray image along with lung field masks prepared using ASM technique

Lung field masks (see Figure 6) are also prepared manually by segmenting the lung fields. Manual segmentation is carried out by determining the peripheral lung field pixel coordinates with segmentation technique.

Figure 6 Original X-ray image and manually segmented lung fields mask



Further, the lung fields are separated from the background by multiplying the mask image with the original filtered X-ray image. The above-discussed technique is applied on SCLC, NSCLC and TB database images.

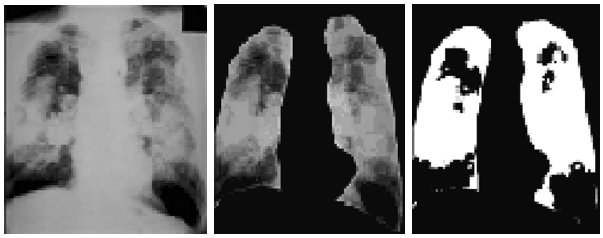
Thresholding is applied on the separated lung field image to separate the nodule or infected portion. Valley point value between the two peaks of the histogram is selected as a threshold value. Region-based segmentation techniques like region-growing (in case of NSCLC) and region-labelling (in case of SCLC) have been applied further to separate the nodule and affected portion.

Figure 7 depicts the results of the nodule segmentation for NSCLC type of image. Results for the SCLC type of image are included in Figure 8.

Figure 7 (a) Segmented lung fields after multiplication; (b) image after thresholding and (c) separated nodule

(a) (b) (c)

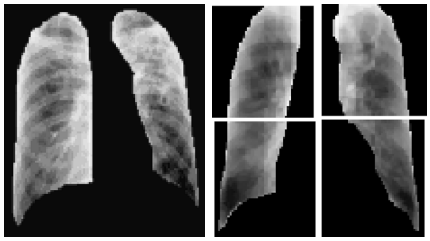
Figure 8 (a) SCLC original image; (b) separated lung fields and (c) separated cancerous portion



(a) (b) (c)

According to the TB database (and Figure 2), TB infection appears mostly in the upper lobes of the lungs. Therefore, separated lung fields image is divided into 4 sections to separate upper and lower lobes (Figure 9).

Figure 9 (a) TB image and (b) divided image



(a) (b)

Bit Quads technique devised by Grey is used to extract geometrical features like *area* and *perimeter*. Distance is a real-valued function $d\{(j_1, k_1), (j_2, k_2)\}$ of two image points (j_1, k_1) (j_2, k_2). The most common measures encountered in image analysis are the Euclidean distance, defined as:

$$d_E = [(j_1 - j_2)^2 + (k_1 - k_2)^2]^{1/2}.$$

In discrete images, the coordinate differences $(j_1 - j_2)$ and $(k_1 - k_2)$ are integers, but the Euclidean distance is usually not an integer. Here, *diameter* is estimated using Euclidean distance. As per the literature survey, it is seen that the growth of the malignant part (nodule over here) is usually circular in nature, therefore roundness of the nodule has been calculated using a simple technique:

$$I = 4 * pi * \text{ area} / \text{perimeter} \wedge 2.$$

This metric value or roundness or circularity index or irregularity index (I) is equal to 1 only for circle and it is less than 1 for any other shapes. Here, it has been assumed that, the more the circularity of the object, the probability of that object being nodule is high.

The geometrical features estimated for the separated nodule, shown in Figure 7(c), are included in Table 1.

Table 1 Geometrical features

S. No.	*Features*	*Value*
1	Area	2815
2	Perimeter	226.85
3	Diameter	59.686
4	Irregularity index (I)	0.69

An important approach for describing a region is to quantify its texture content. A frequently used approach used for texture analysis is based on statistical properties of the intensity histogram. One class of such measures is based on statistical moments. An expression for the nth moment about the mean is given by

$$\mu_n = \sum_{i=0}^{L-1} (z_i - m)^n p(z_i)$$

where z_i is a random variable indicating intensity levels in an image, $p(z)$ is the histogram of the intensity levels in a region, L is the possible intensity levels. A histogram component $p(z_j)$ is an estimate of the probability of occurrence of intensity value z_j, and the histogram may be viewed as an approximation of the Probability Density Function (PDF). GLCM is the technique used to calculate PDF.

$$m = \sum_{i=0}^{L-1} z_i p(z_i)$$

here m is the mean (average) intensity. These moments can be computed using MATLAB function *statmoments*, which is acting as a sub-function in another MATLAB function known as *statxture*. This function is used to calculate first-order statistic texture features like mean, standard deviation, smoothness, third moment, uniformity and entropy.

A measure of average contrast or the standard deviation can be calculated by using the following equation, where $\mu_2(z)$ is the second moment.

$$\sigma = \sqrt{\mu_2(z)} = \sqrt{\sigma^2}.$$

Smoothness measures the relative smoothness of intensity in a region. R is 0 for a region of constant intensity and approaches 1 for region with large excursions in the values of its intensity levels. Smoothness is calculated by using the following equation.

$$R = 1 - 1/(1 + \sigma^2).$$

Skewness of the histogram is also known as third moment. This measure is 0 for symmetric histograms, positive by histograms skewed to the right (about the mean) and negative for histograms skewed to the left. For smooth images, this value comes to be negative. The following equation is used to calculate third moment.

$$\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 \, p(z_i).$$

When all grey levels are equal, uniformity measures maximum and goes on decreasing from there for the inequality.

$$U = \sum_{i=0}^{L-1} p^2(z_i).$$

Entropy is nothing but the measure of randomness, given by the following equation.

$$e = -\sum_{i=0}^{L-1} (pz_i) \log_2 \, p(z_i).$$

The GLCM functions characterise the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix. However, a single GLCM might not be enough to describe the textural features of the input image. For example, a single horizontal offset might not be sensitive to texture with a vertical orientation. Therefore, it is essential to generate multiple GLCMs with different offset values or at different angles. MATLAB function *graycomatrix* is used to generate multiple GLCMs. Using multiple GLCMs contrast or second-order statistic features like Contrast Correlation, Energy and Homogeneity are estimated.

First- and Second-order statistic feature values for the image shown in Figure 8(b) are included in Tables 2 and 3, respectively.

Table 2 Texture features

S. No.	*First-order statistic features*	*Values*
1	Average grey level	48.57
2	Standard deviation	61.36
3	Smoothness	0.06
4	Third moment	2.50
5	Uniformity	0.28
6	Entropy	4.37

Table 3 Texture features

Second-order statistic features	*For offset* [0 1] 0°	*For offset* [–1 1] 45°	*For offset* [–1 0] 90°	*For offset* [–1 –1] 135°	*Avg. value*
Contrast	0.15	0.19	0.11	0.19	0.16
Correlation	0.97	0.97	0.98	0.97	0.97
Energy	0.36	0.35	0.36	0.35	0.35
Homogeneity	0.98	0.97	0.98	0.97	0.97

5 Results and discussion

Being scattered nature of the infection area in case of SCLC type of images (as seen in Figure 8(c)), only *area* of the affected portion is estimated. It is seen that, *irregularity index* in SCLC type of images is always less than 0.1. Geometrical features like *area*, *perimeter* and *irregularity index* have been estimated faithfully for the segmented nodules of NSCLC type of images. First- and second-order statistic features are calculated for all types of images. Substantial difference is seen in the 1st and 2nd order statistic feature values of the upper and lower lobes in case of TB database. Both feature values are marginally larger in upper lobes than the lower one. Results include only upper lobe values. The feature extraction techniques discussed in the above-mentioned section have been applied on 75 images (25 from each category). The results are included only for 10 samples from each category.

Table 4 includes only the *area* values for the SCLC images. Geometrical features like *area*, *perimeter* and *irregularity index* for the NSCLC type of images have been added in Table 5.

Table 4 Geometrical feature

S. No.	*Samples*	*Area*
1	SC-1	35246
2	SC-2	46523
3	SC-3	15252
4	SC-4	14235
5	SC-5	44246
6	SC-6	24729
7	SC-7	18028
8	SC-8	24897
9	SC-9	45551
10	SC-10	54173

Texture-related features or first-order statistic features for SCLC and NSCLC images are shown in Tables 6 and 7, respectively. Second-order statistic features (average values) for SCLC and NSCLC images are included in Tables 8 and 9, respectively. Tables 10 and 11 depict the first- and second-order statistic feature values for TB images.

Table 5 Geometrical features

Samples	*Area*	*Perimeter*	*Irregularity index*
NSC-1	527	97	0.71
NSC-2	347	10	0.55
NSC-3	1015	173	0.42
NSC-4	2377	240	0.52
NSC-5	4098	405	0.31
NSC-6	467	108	0.50
NSC-7	3839	541	0.44
NSC-8	460	111	0.46
NSC-9	1598	175	0.65
NSC-10	3361	264	0.60

Table 6 1st order statistic features for SCLC images

Samples	*Avg. grey level*	*Std. deviation*	*Smoothness*	*Third moment*	*Uniformity*	*Entropy*
SC-1	40.606	65.56	0.062	6.847	0.365	4.079
SC-2	43.878	64.00	0.059	5.067	0.367	3.970
SC-3	11.087	31.29	0.015	1.479	0.733	1.628
SC-4	14.979	46.95	0.032	5.162	0.794	1.344
SC-5	50.151	72.02	0.074	6.904	0.351	4.139
SC-6	24.278	47.47	0.033	3.519	0.525	2.865
SC-7	20.069	45.02	0.030	3.612	0.471	3.166
SC-8	24.479	47.71	0.034	3.536	0.523	2.875
SC-9	48.915	74.46	0.079	7.337	0.419	3.676
SC-10	38.058	55.53	0.045	2.775	0.419	3.449

Table 7 1st order statistic features for NSCLC images

Samples	*Avg. grey level*	*Std. deviation*	*Smoothness*	*Third moment*	*Uniformity*	*Entropy*
NSC-1	30.130	45.015	0.030	1.517	0.437	3.222
NSC-2	7.234	22.829	0.008	0.868	0.535	2.393
NSC-3	30.97	54.85	0.044	4.306	0.502	3.054
NSC-4	17.446	35.498	0.019	1.813	0.395	3.529
NSC-5	11.885	27.354	0.011	1.076	0.298	3.684
NSC-6	13.439	29.925	0.014	1.16	0.386	3.398
NSC-7	27.959	44.359	0.029	1.198	0.446	3.232
NSC-8	7.2056	20.173	0.006	0.559	0.585	2.253
NSC-9	22.145	32.170	0.016	0.738	0.372	3.487
NSC-10	24.083	40.648	0.025	1.515	0.498	2.879

Table 8 2nd order statistic features for SCLC images

Samples	*Contrast*	*Correlation*	*Energy*	*Homogeneity*
SC-1	0.270	0.959	0.409	0.972
SC-2	0.223	0.964	0.384	0.973
SC-3	0.056	0.962	0.736	0.992
SC-4	0.133	0.961	0.792	0.988
SC-5	0.327	0.958	0.366	0.971
SC-6	0.134	0.961	0.534	0.982
SC-7	0.171	0.945	0.619	0.979
SC-8	0.133	0.962	0.532	0.982
SC-9	0.219	0.974	0.432	0.973
SC-10	0.171	0.964	0.441	0.979

Table 9 2nd order statistic features for NSCLC images

Samples	*Contrast*	*Correlation*	*Energy*	*Homogeneity*
NSC-1	0.103	0.965	0.464	0.981
NSC-2	0.072	0.912	0.808	0.986
NSC-3	0.184	0.959	0.509	0.979
NSC-4	0.113	0.944	0.569	0.976
NSC-5	0.081	0.933	0.696	0.982
NSC-6	0.099	0.930	0.664	0.981
NSC-7	0.103	0.966	0.465	0.985
NSC-8	0.072	0.897	0.751	0.986
NSC-9	0.084	0.947	0.428	0.984
NSC-10	0.104	0.959	0.517	0.985

Table 10 1st order statistic features for TB images

Samples	*Avg. grey level*	*Std. deviation*	*Smoothness*	*Third moment*	*Uniformity*	*Entropy*
TB-1	81.72	89.79	0.11	3.17	0.28	4.24
TB-2	96.23	84.28	0.09	0.31	0.16	5.34
TB-3	73.59	78.81	0.09	3.11	0.24	4.76
TB-4	77.19	73.48	0.08	1.69	0.17	5.39
TB-5	68.38	70.83	0.07	2.85	0.19	5.20
TB-6	40.65	52.54	0.04	2.31	0.32	3.94
TB-7	51.57	54.83	0.04	1.62	0.21	4.66
TB-8	50.62	64.73	0.06	3.81	0.30	4.42
TB-9	76.12	74.64	0.06	2.96	0.27	4.98
TB-10	65.98	80.25	0.09	4.96	0.30	4.39

Table 11 2nd order statistic features for TB images

Samples	*Contrast*	*Correlation*	*Energy*	*Homogeneity*
TB-1	0.432	0.972	0.335	0.961
TB-2	0.354	0.946	0.552	0.972
TB-3	0.368	0.968	0.281	0.963
TB-4	0.401	0.965	0.210	0.951
TB-5	0.453	0.962	0.238	0.963
TB-6	0.215	0.963	0.375	0.980
TB-7	0.198	0.965	0.303	0.973
TB-8	0.294	0.958	0.335	0.956
TB-9	0.342	0.968	0.463	0.973
TB-10	0.301	0.962	0.332	0.970

In the image analysis stage, the scattered infected portion (in case of SCLC images) and lung tumour portion (in case of NSCLC images) are separated from the background using segmentation algorithm. In this study, region-based segmentation algorithms (region labelling and region growing) with manual thresholding are adopted to segment the affected portion as well as lung nodules. Since in case of SCLC images, infected portion is more scattered in nature, therefore diameter and perimeter do not provide any meaningful information. Irregularity index is below 0.1 in all the cases. Nodules are separated from the background of the image properly in case of NSCLC images. Circularity is very poor in some of the cases of separated nodules. A diameter value does not provide any meaningful information in such cases. Therefore, geometrical features like area, perimeter and irregularity index have been estimated for the NSCLC images.

To estimate texture-related and contrast features, GLCM technique is used. Most of the times single GLCM may not be sufficient to describe the spatial relationships between the pixels at various directions (or at various angles); therefore, multiple GLCMs are created using various offset values. According to the discussion in features section, TB infection affects mostly the upper lobes of the lungs. Even from the results, it has been observed that, average grey level, standard deviation, entropy and contrast values are larger for the upper two lobes. At the same time, smoothness, uniformity and energy values are seen to be poor for the upper lobes. Therefore, for TB images, texture and contrast-related feature values are estimated only for the upper lobe or lobes (average value), and are included in Tables 10 and 11.

Automatic segmentation is difficult due to the overlapping intensities, anatomic variability in shape, artefacts and noise. If the given image shows large amount of overlapping intensities among various cancer and TB images, then the selection of optimum global threshold becomes difficult and the presented algorithm does not provide satisfactory results. For segmentation of lung tumours from cancer images, the techniques used in this study sets the threshold at valley between two peaks. These algorithms work well for the images where the tumour has brighter contrast when compared with the normal lung tissue. Therefore, knowledge-based techniques can be used for the segmentation of various types of tumours.

In this study, texture and contrast features are extracted using GLCM method. The same or additional features can be extracted by using Discrete Wavelet transform, Duel Trace complex Wavelet Transform to improve the classification accuracy further.

Automatic delineation of the posterior and anterior ribs is a harder problem for which no thoroughly evaluated methods have been proposed yet. The results of texture analysis in chest radiographs are encouraging, but progress is needed to detect more subtle cases.

Segmentation of branching blood vessels and detection of objects such as clothing and catheters has not received much attention, although the results of such analysis could be used to eliminate false positives, to choose ROIs texture analysis and to subtract normal structures. This remains an open problem.

References

Armato, S.G., Giger, M.L. and MacMahon, H. (1998) 'Automated lung segmentation in digitised posterio–anterior chest radiographs', *Academic Rad.*, Vol. 4, pp.245–255.

Brown, M.S., Wilson, L.C., Doust, B.D., Gill, R.W. and Sun, C. (1998) 'Knowledge-based method for segmentation and analysis of lung boundaries in chest X-ray images', *Computerized Med. Imag. Graphics*, Vol. 22, pp.463–477.

Carreira, J., Carrascal, F., Souto, M., Tahoces, P., Gomez, L. and Vidal, J. (1998) 'Automatic calculation of total lung capacity from automatically traced lung boundaries in postero-anterior and lateral digital chest radiographs', *Med Phys.*, Vol. 25, No. 7, pp.1118–1131.

Cootes, T.F. and Taylor, C.J. (1999) *Statistical Models of Appearance for Computer Vision*, Wolfson Image Analysis Unit, Univ. Manchester, Tech. Rep., Manchester, UK.

Duryea, J. and Boone, J.M. (1995) 'A fully automatic algorithm for the segmentation of lung fields in digital chest radiographic images', *Med. Phys.*, Vol. 22, No. 2, pp.183–191.

Giger, M., Doi, K. and MacMahon, H. (1988) 'Image feature analysis and computer-aided diagnosis in digital radiography: automated detection of nodules in peripheral lung fields', *Med Phys.*, Vol. 15, No. 2, pp.158–166.

Giger, M., Doi, K., MacMahon, H., Metz, C. and Yin F. (1990) 'Pulmonary nodules: computer-aided detection in digital chest images', *Radiographics*, Vol. 10, pp.41–51.

Keserci, B. and Yoshida, H. (2002) 'Computerized detection of pulmonary nodules in chest radiographs based on morphological features and wavelet snake model', *Medical Image Analysis*, Vol. 6, pp.431–447.

Li, Q., Katsuragawa, S., Engelmann, R., Armoto, S., MacMahon, H. and Doi, K. (2001) 'Development of a multiple-templates matching technique for removal of false positives in computer-aided diagnostic scheme', *Proc. SPIE*, Vol. 4322, pp.1763–1770.

McNitt-Gray, M.F., Huang, H.K. and Sayre, J.W. (1995) 'Feature selection in the pattern classification problem of digital chest radiograph segmentation', *IEEE Trans. Med. Imag.*, Vol. 14, pp.537–547.

Penedo, M., Carreira, M., Mosquera, A. and Cabello, D. (1998) 'Computer-aided diagnosis: a neural-network-based approach to lung nodule detection', *IEEE Trans. Med. Imag.*, Vol. 17, pp.872–880.

Pietka, E. (1994) 'Lung segmentation in digital chest radiographs', *J. Digital Imag.*, Vol. 2, pp.79–84.

Sankar P. and Sklansky J. (1982) 'A Gestalt guided heuristic boundary follower for X-ray images of lung nodules', *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. PAMI – 4, pp.326–331.

Sklansky, J. and Petkovic, D. (1984) 'Two-resolution detection of lung tumors in chest radiographs', in Rosenfeld, A. (Ed.): *Multiresolution Image Processing and Analysis*, Springer-Verlag, Berlin, Germany, pp.365–378.

Toriwaki, J., Ballard, D.H. and Lawpeter, W.A. (1973) 'Pattern recognition of chest x-ray images', *Computer Graphics and Image Processing*, Vol. 2, pp.375–390.

Tsujii, O., Freedman, M.T. and Mun, S.K. (1998) 'Automated segmentation of anatomic regions in chest radiographs using an adaptive-sized hybrid neural network', *Med. Phys.*, Vol. 25, No. 6, pp.998–1007.

Van Ginneken, B., Katsuragawa, S., ter Haar Romeny, B.M., Doi, K. and Viergever, M.A. (2002) 'Automatic detection of abnormalities in chest radiographs using local texture analysis', *IEEE Transactions on Medical Imaging*, Vol. 21, No. 2, pp.139–149.

Van Ginneken, B. and ter Haar Romeny, B.M. (1995) 'Automatic segmentation of lung fields in chest radiographs', *Med. Phys.*, Vol. 27, No. 10, pp.2445–2455.

Vittitoe, N.F., Vargas-Voracek, R. and Floyd Jr., C.E. (1998) 'Identification of lung regions in chest radiographs using Markov Random Field modeling', *Med. Phys.*, Vol. 25, No. 6, pp.976–985.

Wei, J., Hagihara, Y., Shimizu, A. and Kobatake, H. (2002) *Optimal Image Feature Set for Detecting Lung Nodules on Chest X-Ray Images*, CARS/Springer, New York.

Xu, X.W. and Doi, K. (1995) 'Image feature analysis for computer-aided diagnosis: accurate determination of ribcage boundary in chest radiographs', *Med. Phys.*, Vol. 22, No. 5, pp.617–626.

Yue, Z., Goshtasby, A. and Ackerman, L. (1995) 'Automatic detection of rib borders in chest radiographs', *IEEE Trans. Med. Imag.*, Vol. 14, pp.525–536.

Zhang, J. and Huang, H.K. (1997) 'Automated lung segmentation in x-ray computed tomography', *Acad. Radiol*., Vol. 10, No. 11, pp.1221–1230.

Bibliography

Behiels, G., Vandermeulen, D., Maes, F., Suetens, P. and Dewaele, P. (1999) *Active Shape Model-Based Segmentation of Digital X-Ray Images*, Lecture Notes in Computer Science, MICCAI '99, Springer-Verlag, Berlin, Germany, Vol. 1679, pp.128–137.

Cootes, T.F., Taylor, C.J., Cooper, D. and Graham, J. (1995) 'Active shape models: their training and application', *Computer Vis. Image Understanding*, Vol. 61, No. 1, pp.38–59.

Dryden, I. and Mardia, K.V. (1998) *The Statistical Analysis of Shape*, Wiley, London, UK.

Ginneken, B.V. (2001) *Computer Aided Diagnosis in Chest Radiography*, PhD Thesis, Netherlands.

Gonzalez, R.C., Woods, R.E. and Eddins, S.L. (2002) *Digital Image Processing*, 2nd ed., Pearson Education, Delhi.

Kelemen, A., Székely, G. and Gerig, G. (1999) 'Elastic model-based segmentation of 3D neuroradiological data sets', *IEEE Trans. Med. Imag.*, Vol. 18, pp.828–839.

Kupinski, M., Giger, M.L., Lu, P. and Huo, Z. (1995) 'Computerized detection of mammographic lesions: performance of artificial neural network with enhanced feature extraction', *Proc. SPIE*, Vol. 2434, pp.598–605.

Kupinski, M.A. and Giger, M.L. (1997) *Feature Selection and Classifiers for the Computerized Detection of Mass Lesions in Digital Mammography*, IEEE International Congress on Neural Networks, Houston, Texas, June, pp.2460–2463.

Lin, J., Lo, S.B., Hasegawa, A., Freedman, M.T. and Mun, S.K. (1996) 'Reduction of false positives in lung nodule detection using a two-level neural classification', *IEEE Trans. Med. Imag.*, Vol. 15, pp.206–217.

Pratt, W.K. (2002) *Digital Image Processing*, 3rd ed., A Wiley – Interscience Publication, Singapore.

Sahiner, B., Chan, H.P., Wei, D., Petrick, N., Helvie, M.A., Adler, D.D. and Goodsitt, M.M. (1996) 'Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue', *Med. Phys.*, Vol. 23, No. 10, pp.1671–1683.

Schilham, A.M.R., van Ginneken, B. and Loog, M. (2006) 'A computer aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database', *Med. Img. Analysis*, Vol. 10, pp.247–258.

Suzuki, H., Inaoka, N., Takabatake, H., Mori, M., Natori, H. and Suzuki, A. (1991) 'An experiment system for detecting lung nodules by chest X-ray image processing', *SPIE. Biomedical Image Processing II*, Vol. 1450, pp.99–107.

Tourassi, G.D., Frederick, E.D., Markey, M.K. and Floyd, C.E. (2001) 'Application of the mutual information criterion for feature selection in computer-aided diagnosis', *Med. Phys.*, Vol. 28, No. 12, pp.2394–2402.

Van Ginneken, B., ter Haar Romeny, B.M. and Viergever, M.A. (2001) 'Computer-aided diagnosis in chest radiography: a survey', *IEEE Trans. Med. Imag*, Vol. 20, No. 12, pp.1228–1230.

Wu, Y., Giger, M.L., Doi, K., Vyboorny, C.J., Schmidt, R.A. and Metz, C.E. (1993) 'Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer', *Radiology*, Vol. 187, pp.81–87.

Xu, X-W., Doi, K., Kobayashi, T., MacMahon, H. and Giger, M.L. (1997) 'Development of an improved CAD scheme for automated detection of lung nodules in digital chest images', *Med. Phys.*, Vol. 24, No. 9, pp.1395–1403.

Yoshida, H. and Doi, K. (2000) 'Computerized detection of pulmonary nodules in chest radiographs: reduction of false positives based on symmetry between left and right lungs', *Proc. SPIE in Medical Imaging 2000*, pp.97–102.