Journal of Scientific & Industrial Research
Vol. 69, April 2010, pp. 271-277

Chest X-ray features extraction for lung cancer classification

S A Patil* and V R Udupi
Textile & Engineering Institute, Ichalkaranji 416 115, India

Received 23 April 2009; revised 11 January 2010; accepted 13 January 2010

This study presents a computer algorithm, which consists of four main steps (image acquisition, image pre-processing, nodule candidate detection, and feature extraction) for nodule detection in chest radiographs. Algorithm is applied on small-cell type of lung cancer (SCLC) and non-small-cell type of lung cancer (NSCLC) images. Total 50 images (25 from each category) were used to estimate geometrical and texture features. Active shape model (ASM) was used for lung field segmentation. Gray level co-occurrence matrix (GLCM) was used to estimate texture features.

Keywords: Active shape model (ASM), Chest X-ray, Gray level co-occurrence matrix (GLCM), Lung field segmentation

Introduction

Prognosis and cure of lung cancer depend highly on early detection and treatment of small and localized tumors. Survival rate of a patient (5 y old cancer) is approx. 40% when lung cancer is detected in early stage. Lung cancers (87%) are thought to result from smoking or passive smoking. Physical characteristics of nodules (rate of growth, pattern of calcification, type of margins) are very important in investigation of solitary lung nodules. Malignant nodule's have a fast doubling time (25-450 days), whereas benign nodules are stable (doubling time, > 500 days)[1]. Computer-aided diagnosis (CAD) is a very effective approach for improving diagnostic accuracy. Numerous systems are reported for detecting lung nodules on chest X-ray images[2-4]. To reduce number of false positives while maintaining a high true positive detection rate is the most important work in realizing a chest CAD system[5]. There have been a few studies[6-10] on comparing effectiveness of features proposed for discriminating between normal and abnormal tissues.

This study presents optimal feature set for classification from available database.

Experimental

Features

As a cancer develops, cancer cells may produce chemicals that cause to form nearby new blood vessels, which nourish cancer cells to grow and form a tumor large enough to see on X-rays. Cells from cancer can break away from original tumor and spread to other body parts. Lung cancers are classified as small-cell type of lung cancer (SCLC), and non-small-cell type of lung cancer (NSCLC) (Fig. 1). Usually SCLC arises at alveolar level or at terminal bronchial level, and seen to be more scattered on X-ray. NSCLC arises in larger, more central bronchi; tends to spread locally; and metastasizes somewhat larger than other patterns, but its rate of growth in its site of origin is usually more rapidly than that of other types.

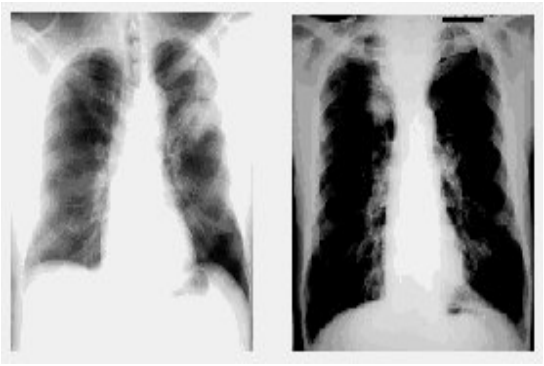

Fig. 1—SCLC and NSCLC images

Feature selection is a very important step in organizing a classifier. Features are classified[11] as, geometric, contrast, first order statistics, and second order statistics features. Toriwaki *et al*[12] detected suspected nodule areas (SNA) using image processing technology[12], and could only detect nodules from approx.

*Author for correspondence
Telefax: +91 230 – 2432329
E-mail: shrinivasapatil@gmail.com

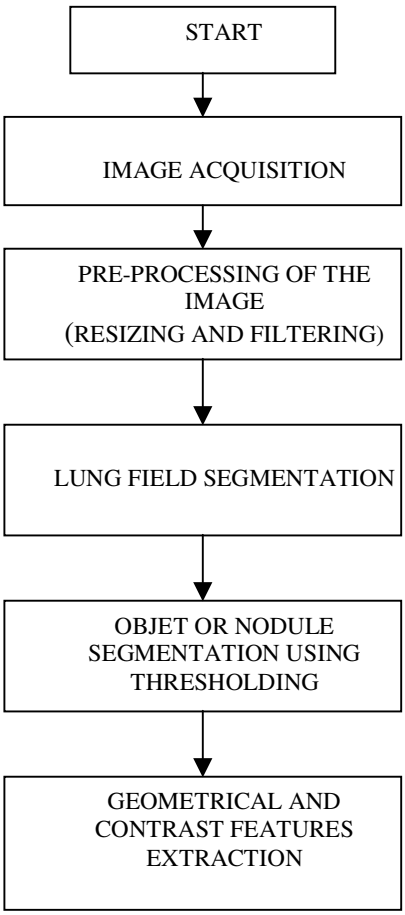

Fig. 2—Feature extraction technique

1 cm. In recent research, several methods have been proposed[1] to reduce number of false positives (FP's) while maintaining high sensitivity. Some morphology-based algorithms[1] have been proposed to extract specific features (circularity, size, contrast, or local curvature). During 1995, an algorithm was developed for detection of posterior rib borders, which is one of the FP's in chest radiographs. Automatic background recognition and removal (ABRR) method[1] for chest radiographs has shown an excellent performance (99%) of correct recognition and removal of (91%) background signal. During 1998, CAD system was developed to detect big nodules (not better for small nodules) when in initial stage. During 2002, a fully automatic scheme was presented for texture analysis of lung fields in chest radiographs.

Method

Images are obtained using image acquisition method and then applied pre-processing algorithms, including size normalization and image filtering (Fig. 2). Features, useful for diagnosis and analysis, require separation of lung fields from background. Lung field masks, prepared manually by segmenting lung fields, as well as readily available masks developed by using ASM technique[13], were used to separate lung fields (JSRT public database). Thresholding along with region based segmentation techniques were used to segment lung nodules (in NSCLC images) and cancerous portion (in SCLC images) from separated lung field area. In next step, geometrical and contrast features were estimated. Chest unit used for screening X-ray films was mobile KlinoskopH unit (Siemens, India). Keeping tube voltage equal to 150 kV, 500 mA at 2.2 mm Pb, images were printed on film (14 cm X 17 cm), which was digitized with a high-resolution scanner (Scanjet 2400, HP India). SCLC as well as NSCLC images from public database were also used in this study. Every image data was acquired with 256 gray levels (8 bits) and stored as JPEG (.jpg, .jpeg) data. Before extraction of features, image is pre-processed to reduce irrelevant information or noise, and to enhance image properties, which makes feature measurement easier and more reliable. Scanned images were resized (512 X 512 pixels) and, then Median filter was used to remove noise or irrelevant information from images.

Segmentation of lung fields on PA chest radiographs has received considerable attention[14-19]. Lung segmentation by pixel classification using neural networks has been investigated[20,21]. Vittitoe *et al*[22,23] developed a pixel classifier for identification of lung regions using Markov random field modeling. Ginneken & Romeny[24] proposed a hybrid method [improved Active Shape Modeling (ASM) technique] that combines a rule-based scheme with a pixel classifier. ASM[25-27] has been applied to various segmentation tasks in medical imaging. Under ASM scheme, an object is described by landmark points, which are (manually) determined in a set of training images. From collections of landmark points, a point distribution model is constructed[28, 29].

To create models of image profiles around each landmark, profiles ($g_1 \ldots g_n$) are sampled around each landmark, perpendicular to the contour. Sampling k pixels on either side of profiles gives profiles of length, $2k + 1$. First derivatives of these profiles are used. Profiles are also normalized by dividing through the sum of absolute values of elements. For each landmark point, mean profile g' and covariance matrix Sg are computed. To fit model,

Mahalanobis distance between a new profile g_i and profile model can be computed as

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

Minimizing Mahalanobis distance $f(g_i)$ is equivalent to maximizing probability that g_i originates from distribution $\{g_1 \ldots g_n\}$. This minimization is used to find new locations for landmarks during fitting. These profile models, given by g' and Sg', are constructed for multiple resolutions. Finest resolution uses original image and a step size (1 pixel) when sampling. Next resolution is image observed at scale $\sigma = 1$ and step size of 2 pixels. Subsequent levels are constructed by doubling image scale and step size. Shapes are fitted in an iterative manner, starting from mean shape. Each landmark is moved along direction to the contour to n_s positions on either side, evaluating a total of $2n_s + 1$ position. Landmark is put at the position with lowest Mahalanobis distance. After moving all landmarks, shape model is fitted to the points (Fig. 3), yielding an updated segmentation. This is repeated *N* times at each resolution, from coarse to fine. Around 180 lung field masks (Fig. 4) prepared using ASM technique are available (JSRT and SCR public database) for lung field segmentation (separately for each right and left lung fields).

Lung field masks (Fig. 5) are also prepared manually by segmentation technique, which is carried out by determining peripheral lung field pixel co-ordinates with segmentation technique. Further, lung fields are separated from background by multiplying mask image with original X-ray image. Thresholding is applied on separated lung field image to separate nodule or cancerous portion. Valley point value between two peaks of histogram is selected as a threshold value. Region based segmentation techniques[30] like region-growing [NSCLC (Fig. 6)], and region-labeling [SCLC (Fig. 7)] were applied further to separate nodule and affected portion.

Bit Quads technique[31] was 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). Most common measures encountered in image analysis are Euclidean distance, defined as

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

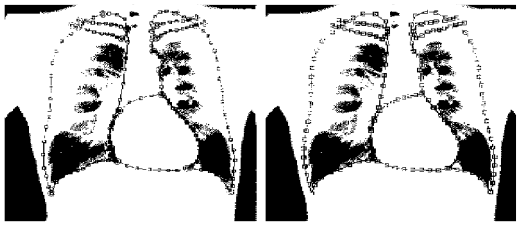

Fig. 3—Landmark points with lung fields

Fig. 4—Original X-ray image along with lung field masks prepared using ASM technique

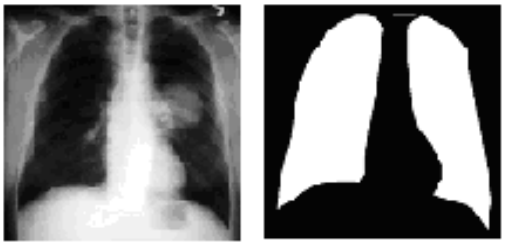

Fig. 5—Original X-ray image and manually segmented lung fields mask

Fig. 6— (a) Segmented lung fields after multiplication, (b) Image after thresholding, (c) Separated nodule

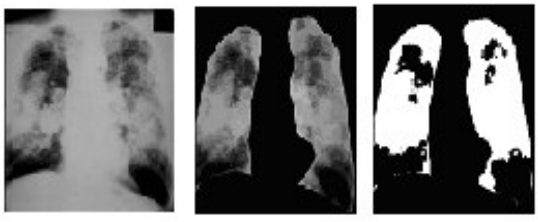

Fig. 7—(a) SCLC original image, (b) Separated lung fields, (c) Separated cancerous portion

In discrete images, coordinate differences $(j_1 - j_2)$ and $(k_1 - k_2)$ are integers, but Euclidean distance is usually not an integer. Here *diameter* is estimated using Euclidean distance. Growth of malignant part

(nodule over here) is usually circular in nature, therefore roundness of nodule has been calculated as

$$I = 4 * \text{pi} * \text{area} / \text{perimeter} \wedge 2 \quad ...(3)$$

This metric value or roundness or circularity index or irregularity index (I) is equal to 1 only for circle and it is < 1 for any other shape. Here it has been assumed that, more circularity of the object, the probability of that object being nodule is high. Geometrical features estimated for separated nodule (Fig. 6c) has been found as follows: area, 2815; perimeter, 226.85; diameter, 59.686; and I, 0.69. A frequently used approach for texture analysis is based on statistical properties of intensity histogram. One class of such measures is based on statistical moments. An expression for nth moment about mean is given as

$$\mu_n = \sum_{i=0}^{L-1} (z_i - m)^n p(z_i) \quad ...(4)$$

where z_i is a random variable indicating intensity levels in an image, $p(z)$ is histogram of intensity levels in a region, L is possible intensity levels. A histogram component, $p(z_j)$, is an estimate of probability of occurrence of intensity value, z_j, and histogram may be viewed as an approximation of probability density function (PDF). Gray level co-occurrence matrix (GLCM) technique is used to calculate PDF. Mean (average) intensity is calculated as

$$m = \sum_{i=0}^{L-1} z_i p(z_i) \quad ...(5)$$

These moments can be computed using MATLAB function *statmoments*, which acts as a sub-function in another MATLAB function known as *statxture*[32]. This function is used to calculate first-order statistic texture features (mean, standard deviation, smoothness, third moment, uniformity and entropy). Average contrast or standard deviation can be calculated as

$$\sigma = \sqrt{\mu_2(z)} = \sqrt{\sigma^2} \quad ...(6)$$

where $\mu_2(z)$ is second moment.

Smoothness measures relative smoothness of intensity in a region. R is 0 for a region of constant intensity and approaches 1 for region with large excursions. Smoothness is calculated as

$$R = 1 - 1/(1 + \sigma^2) \quad ...(7)$$

Skewness of histogram or third moment 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. Third moment is calculated as

$$\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i) \quad ...(8)$$

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

$$U = \sum_{i=0}^{L-1} p^2(z_i) \quad ...(9)$$

Entropy, measure of randomness, is given as

$$e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i) \quad ...(10)$$

GLCM functions characterize 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 textural features of 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[32] *graycomatrix* is used to generate multiple GLCMs. Using multiple GLCMs, contrast or second-order statistic features (contrast correlation, energy, and homogeneity) are estimated.

First order statistic features (Fig. 7c) are as follows: average gray level, 48.57; standard deviation, 61.36; smoothness, 0.06; third moment, 2.50; uniformity, 0.28; and entropy, 4.37. Second order

Table 1—Texture features

Second-order statistic features	For offset [0 1] 0°	[-1 1] 45°	[-1 0] 90°	[-1-1] 135°	Average 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

Table 2—Geometrical features for SCLC an NSCLC images

Samples	SCLC images	NSCLC images		
	Area	Area	Perimeter	Irregularity index
1	35246	527	97	0.71
2	46523	347	10	0.55
3	15252	1015	173	0.42
4	14235	2377	240	0.52
5	44246	4098	405	0.31
6	24729	467	108	0.50
7	18028	3839	541	0.44
8	24897	460	111	0.46
9	45551	1598	175	0.65
10	54173	3361	264	0.60

statistic feature values for image (Fig. 7c) are included in Table 1.

Results and Discussion

Being scattered nature of cancerous portion in SCLC type of images (Fig. 7b), only *area* of affected portion was estimated. In SCLC type of images, *I* is always < 0.1. Geometrical features were estimated for NSCLC type of images. First and second order statistic features are calculated for both types of images. Feature extraction techniques have been applied on 50 images (25 from each category). Results are included only for 10 samples from each category. Geometrical features (area, perimeter, and *I*) are included for NSCLC type of images (Table 2) and only *area* values for SCLC images. Texture related features or first-order statistic features are calculated for SCLC (Table 3) and NSCLC (Table 4) images. Second-order statistic features (average values) are calculated for SCLC (Table 5) and NSCLC (Table 6) images. In the next phase, discriminating features of lung cancer images are selected for classification purpose. ANN based backpropagation technique was

Table 3—1st order statistic features for SCLC images

Samples	Average gray level	Standard 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 4—1st order statistic features for NSCLC images

Samples	Average gray level	Standard 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 5—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 6—2nd order texture 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

used to classify lung cancers. Texture features related to normal lung images were also considered during classification.

Conclusions

In case of SCLC images, area values were found quite larger than other counterpart because of scattered nature of affected portion. Irregularity index is always closer to '1' for circular objects. Segmented portion in NSCLC images are having irregularity index closer to '1', indicating segmented portion is a malignant portion or a lung nodule. Uniformity and energy values are almost identical. Among other features (standard deviation, average gray level, smoothness, third moment, entropy, and contrast), values are more dominating (or larger) in SCLC images than NSCLC images due to scattered nature of affected portion throughout the lung fields. Features like correlation and homogeneity are almost identical in both the cases.

References

1 van Ginneken B, ter Haar Romeny B M & Viergever M A, Member IEEE, Computer-aided diagnosis in chest radiography: a survey, *IEEE Trans Med Imag*, **20** (2001)1228-1230.

2 Suzuki H, Inaoka N, Takabatake H, Mori M, Natori H & A Suzuki, An experiment system for detecting lung nodules by chest x-ray image processing, *SPIE Biomed Image Process II*, **1450** (1991) 99-107.

3 Lin J, Lo S B, Hasegawa A, Freedman M T & Mun S K, Reduction of false positives in lung nodule detection using a two-level neural classification, *IEEE Trans Med Imag*, **15** (1996) 206-217.

4 Xin-Wei. Xu, Doi K, Kobayashi T, MacMahon H & Giger M L, Development of an improved CAD scheme for automated detection of lung nodules in digital chest images, *Med Phys,* **24** (1997) 1395-1403.

5 Yoshida H & Doi K, Computerized detection of pulmonary nodules in chest radiographs: reduction of false positives based on symmetry between left and right lungs, *Proc SPIE Med Imag,* **2000** (2000) 97-102.

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

7 Kupinski M, Giger M L, Lu P & Huo Z, Computerized detection of Mammographic lesions: Performance of artificial neural network with enhanced feature extraction, *Proc SPIE,* **2434** (1995) 598-605.

8 Sahiner B, Chan H P, Wei D, Petrick N, Helvie M A, Adler D D & Goodsitt M M, Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue, *Med Phys,* **23** (1996) 1671-1683.

9 Kupinski M A & Giger M L, Feature selection and classifiers for the computerized detection of mass lesions in digital mammography, *IEEE Int Congr on Neural Networks* (Houston, Texas) 1997, 2460-2463.

10 Tourassi G D, Frederick E D, Markey M K & Floyd C E, Application of the mutual information criterion for feature selection in computer-aided diagnosis, *Med Phys,* **28** (2001) 2394-2402.

11 Wei J, Hagihara Y, Shimizu A & Kobatake H, *Optimal Image Feature Set for Detecting Lung Nodules on Chest X-Ray Images*, **2nd edn** (CARS/Springer, Berlin, Ggermany) 2002, 139-210.

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

13 van Ginneken B, *Computer aided diagnosis in chest radiography*, Ph D Thesis, Netherland Institute of Technology, Neetherland, 2001.

14 Armato S G, Giger M L & MacMahon H, Automated lung Segmentation in digitized postero–anterior chest radiographs, *Acad Rad*, **4** (1998) 245-255.

15 Xu X W & Doi K, Image feature analysis for computer-aided diagnosis: Accurate determination of ribcage boundary in chest radiographs, *Med Phys*, **22** (1995) 617-626.

16 Arnold M R & Mackmohan H, Image feature analysis for computer-aided diagnosis: Detection of right and left hemi-diaphragm edges and delineation of lung field in chest radiographs, *Med Phys*, **23** (1996) 1613-1624.

17 Duryea J & Boone J M, A fully automatic algorithm for the segmentation of lung fields in digital chest radiographic images, *Med Phys*, **22** (1995) 183-191.

18 Pietka E, Lung segmentation in digital chest radiographs, *J Digital Imag*, **2** (1994) 79-84.

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

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

21 Tsujii O, Freedman M T & Mun S K, Automated segmentation of anatomic regions in chest radiographs using an adaptive-sized hybrid neural network, *Med Phys*, **25** (1998) 998-1007.

22 Vittitoe N F, Vargas-Voracek R & Floyd C E, Identification of lung regions in chest radiographs using Markov random field modeling, *Med Phys*, **25** (1998) 976-985.

23 Floyd C E, Markov random field modeling in posteroanterior chest radiograph segmentation, *Med Phys*, **26** (1999) 1670-1677.

24 Ginneken van B & ter Haar Romeny B M, Automatic segmentation of lung fields in chest radiographs, *Med Phys*, **27** (2000) 2445-2455.

25 Cootes T F, Taylor C J, Cooper D & Graham J, Active shape models - Their training and application, *Compu Vis Image Understanding*, **61** (1995) 38-59.

26 Behiels G, Vandermeulen D, Maes F, Suetens P & Dewaele P, Active shape model- based segmentation of digital X-ray images, in *Lecture Notes in Compu Sci*, **vol 1679** (Springer-Verlag, Berlin Germany) (1999) 128-137.

27 Kelemen A, Székely G & Gerig G, Elastic model-based segmentation of 3-d neuroradiological data sets, *IEEE Trans Med Imag*, **18** (1999) 828-839.

28 Cootes T F & Taylor C J, *Statistical models of appearance for computer vision*, **Tech Rep** (Wolfson Image Analysis Unit, Manchester Univ, Manchester, U K) 1999.

29 Dryden L & Mardia K V, *The Statistical Analysis of Shape*, **5th edn** (Wiley, London) 1998.

30 Pratt W K, *Digital Image Processin*, **3rd edn** (Wiley-Interscience Publication, Singapore) 2002.

31 Gonzalez R C, Woods R E & Eddins S L, *Digital Image Processing,* **2nd edn** (Pearson Education, New Delhi) 2002.

32 Gonzalez R C, Woods R E & Eddins S L, *Digital Image Processing Using MATLAB,* **LP edn** (Pearson Education, Delhi) 2004.