Detection of Lung Nodules from CT Images using Image Processing Techniques

Swathi H K* and Nanda S**
*Final Year, M.Tech - BMSP&I, Dept. of IT, SJCE, Mysuru
swathihk1993@gmail.com
*Assistant Professor, Dept. of IT, SJCE, Mysuru
nanda_prabhu@sjce.ac.in

Abstract: Lung cancer is the common cause of death among people throughout the world. So, early detection of lung cancer can increase the chance of survival among the people. The overall 5-year survival rate for lung cancer patients increases from 14 to 49% if the disease is detected in time. Computed Tomography (CT) can be more efficient than X-ray in detecting the lung cancer. In this paper, an algorithm is developed for the segmentation of lung from CT images using optimal thresholding technique. Then Vector quantization (VQ) is performed on Lung CT images to segment lung nodule. In vector Quantization codebook is generated for each image and the training process is done according to the input data. The performance evaluation of segmentation is done using measures like Area, Sensitivity, DICE co-efficient, Jaccard Index, Hausdroff distance and Mahalanobis distance. Texture features, , morphological features and wavelet based features are extracted from the segmented nodule. K-nearest neighbor (KNN), Support Vector Machine and Decision tree are used for the classification of lung nodules as benign or malignant. The performances of these classifiers are compared.

Keywords: Computed tomography images, Optimal thresholding, Texture features, Vector quantization (VQ).

Introduction

Lung cancer is the leading cause of cancer deaths in the world [1]. In developed countries, patients diagnosed with this pathology have a 5 year survival rate between 10 and 16%. This occurs because about 70% of lung cancer cases are diagnosed in advanced stages, preventing effective treatments. However, in cases where lung cancer is diagnosed in early stages, the 5 year survival rate increases to 70% [2]. Earlier chest X-ray was the effective method for the identification of the lung nodule, because of its low contrast image, detection of nodules becomes difficult for the radiologists. Computerized tomography (CT) has become the most sensitive imaging modality for the detection of small lung nodules, particularly because of its helical multi-slice technology [3]. More recently, one of the hopes to change the scenario of late diagnosis has been conducted by monitoring programs with low-dose CT, particularly applied to risk groups such as smokers [4]. Segmentation based on watershed and thresholding methods are reported in [5][6]. In [7] used decimated wavelet transform along with k-means clustering for segmentation of lung nodules in CT images. In [8] the paper proposed a Computer aided diagnosis (CAD) system for lung cancer detection in CT images using histogram thresholding. In [9] has developed a hierarchical vector quantization (VQ) approach to overcome the disadvantages thresholding methods. In [10] have come out with an approach for the detection of components in brain MRI using VQ. From the literature it is clear that VQ is the efficient method for segmentation. In this work the lung nodules are segmented from CT images. Texture features, morphological features and wavelet based features are extracted from the segmented nodule. KNN, SVM and decision tree are used to classify the lung nodule as benign or malignant.

In this paper the method involved in the algorithm is discussed in the following sequence: pre-processing, segmentation, feature extraction and classification. The experimental results and the conclusions are discussed in the final sections.

Methodology

This work involves the methods in the following sequence: pre-processing, segmentation, feature extraction and classification. The block diagram of the work is as shown in Fig.1.

Pre-processing

The Input Lung CT image is filtered using a median filter of size 5x5 in order to remove noise and smoothen the image. The average Peak Signal to Noise Ratio of the filtered image is 18.22dB. The filtered image is shown in Fig.2.

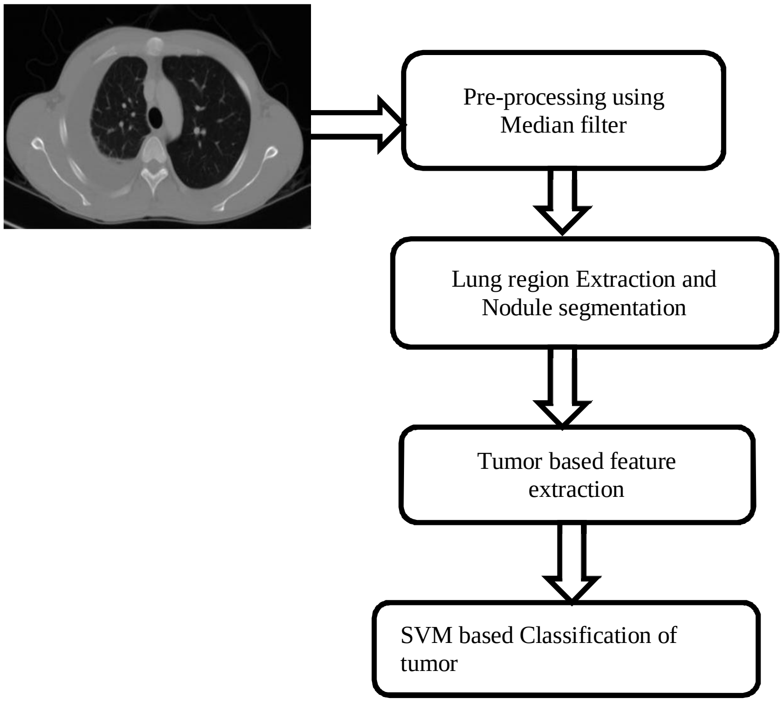
Pre-processing using Median filter
Lung region Extraction and Nodule segmentation
Tumor based feature extraction
SVM based Classification of tumor

Fig. 1: Block diagram of the method

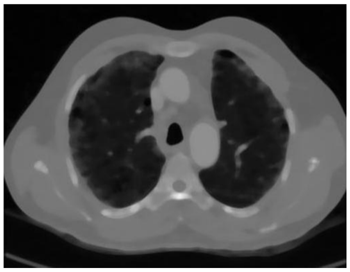

Fig. 2**.** Median Filtered Image

Segmentation

The segmentation of Lung nodules from the CT images fall into two categories; namely, segmentation of lung region from the CT image using Optimal thresholding and extraction of nodule from the lung region using Vector quantization (VQ).

Optimal thresholding

The Optimal thresholding technique is used to identify the regions based on their respective density distribution. Instead of using fixed threshold value to segment the lung region, a histogram analysis method that determines dynamic threshold value for the CT slice automatically. Histogram of an image provides a scheme about probability density values of pixels, these values are derived from the histogram to model the segmentation of lung from CT slice. The result of the segmentation of lung is as shown in Fig.3.

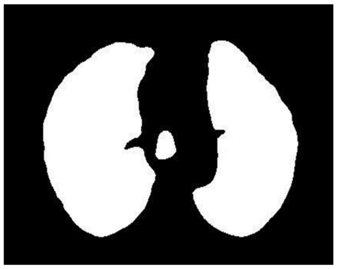

Figure.3. Segmentation of lung using Optimal thresholding

Vector quantization (VQ)

Vector quantization (VQ) is a technique in which a codebook is generated for each image. A codebook is a representation of the entire image containing a definite pixel pattern which is computed according to a specific VQ algorithm. The image is divided into fixed sized blocks that form the training vector. The generation of the training vector is the first step to cluster formation, on these training vectors clustering methods is applied and codebook is generated. The method most commonly used to generate codebook is the Linde-Buzo-Gray (LBG) algorithm which is also called as Generalized Lloyd Algorithm (GLA). Initially, selected 8 as codebook. Thus the image is divided into 8 clusters, thus obtained were mapped onto the image generating 8 different images representing them. All these images were superimposed on the original image giving clear demarcation of the tumor. The nodule segmented is as shown in Fig.4.

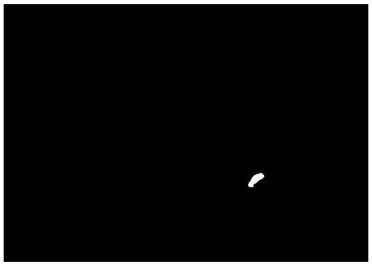

Figure.4 Segmented lung nodule using VQ

Once the nodules are segmented, region based measure are used for validating segmentation results. The measures are Area, Sensitivity, DICE co-efficient, Jaccard Index, Hausdroff distance and Mahalanobis distance.

Area Error Rate estimates difference between occupied areas and is used to quantitatively assess the segmentation accuracy,

$$\text{Area} = \frac{a_{VR} - a_{IR}}{a_{MSR}} \tag{1}$$

where a_{VR} represents the number of pixels in the original image and a_{IR} represents the is the amount of pixels in segmented image, a_{MSR} is the number of pixels in the manually extracted area.

DICE co-efficient is used to compare similarities of two sample and is calculated as,

$$\text{DICE} = \frac{2 \times a_{IR}}{a_{VR} + a_{MSR}} \tag{2}$$

Sensitivity is calculated using

$$\text{Sensitivity} = \frac{a_{IR}}{a_{MSR}} \tag{3}$$

Handsroff distance is a measure for the dissimilarities of two shapes and is used here for comparing the boundary detected. It is given by,

$$\text{HD} = \max[\min\|Cw - Cx\|) \tag{4}$$

Where Cw is the boundary of ASR and C_x is boundary of MSR.

Jaccard distance, which measures *dis*similarity between sample sets, and is calculated as

$$\text{Jaccard index} = 1\text{-DICE} \tag{5}$$

Mahalanobis distance measures the dissimilarities of two images that is, original image and segmented image ,is calculated as,

$$\text{MD} = \text{sqrt}((x-y)C^{-1}(x-y)) \tag{6}$$

C stand for the covariance function, the new (Mahalanobis) distance between two points x and y is the distance from x to y divided by the square root of $C(x-y,x-y)$

Feature Extraction

Feature extraction plays an important role in the characterization of the lung nodule. In this work a total of 15 features are extracted. 11 textural features and 4 wavelet based features.

Haralick's Textural Feature Extraction

Texture is an important characteristic used in identifying objects or regions in an image [11]. The texture information is adequately specified in a set of gray-tone spatial-dependency matrices (Gray Level Co-occurrence Matrix (GLCM) [11]) which are computed for various angular relationships between neighboring cell pairs in the image. From the matrix, a number of textural features of a given image can be computed. Following features are extracted from the segmented lung nodule.

$$\text{Contrast} = \sum_{i=1}^{N-1} P(i,j)(i-j)^2 \tag{7}$$

$$\text{Energy} = \sum_{i=1}^{M} P(i,j)^2 \tag{8}$$

$$\text{Variance} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[P(i,j)-\mu]^2 \tag{9}$$

$$\text{Mean} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} P(i,j) \tag{10}$$

$$\text{Entropy} = \sum_{i=1}^{M}\sum_{j=1}^{N} P(i,j)\log(p(i,j)) \tag{11}$$

$$\text{IDM} = \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{p(i,j)}{i+|i-j|} \tag{12}$$

$$\text{Kurtosis} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left[\frac{[P(i,j)-\mu]}{\sigma}\right]^4 \tag{13}$$

$$\text{Skewness} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left[\frac{[P(i,j)-\mu]}{\sigma}\right]^3 \tag{15}$$

$$\text{Smoothness} = 1-\frac{1}{1+\sigma^2} \tag{16}$$

$$\text{Correlation} = \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{P(i,j)(j-\mu_i)(i-\mu_j)}{\sigma_i\sigma_j} \tag{17}$$

$$\text{Eccentricity} = \frac{(\mu_i-\mu_i)^2+4\times\mu_{i,j}}{\text{Area}} \tag{19}$$

Wavelet based feature Extraction

Discrete wavelet transform decomposes the image into approximations using low-pass and high-pass filtering. Each image may be represented as a $m\times n$ gray-scale matrix $P[i,j]$, where each element of the matrix represents the grayscale intensity of one pixel of the image. Eight different pixels surround and form the neighboring pixels elements of the matrix. In this work db8 is used

The first level of wavelet decomposition yields four co-efficient matrices, namely, *Da*1, *Dh*1, *Dv*1, and *Dd1*. The following features are calculated from these matrices.

$$\text{Average Dh1} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|Dh1(i,j)| \tag{20}$$

$$\text{Average Dv1} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|Dv1(i,j)| \tag{21}$$

$$\text{Energy1} = \frac{1}{M^2 * N^2}\sum_{i=1}^{M}\sum_{j=1}^{N}(Dv1(i,j))^2 \tag{22}$$

$$\text{Energy2} = \frac{1}{M^2 * N^2}\sum_{i=1}^{M}\sum_{j=1}^{N}(Dh1(i,j))^2 \tag{22}$$

Classification

The classification of the nodules as benign and malignant is performed using K-NN, Support Vector Machine with linear, polynomial and RBF kernel and Decision tree. The **k-nearest-neighbor classifier** (**KNNC** for short) is one of the most basic classifiers for pattern recognition or data classification. The principle of this method is based on the intuitive concept that data instances of the same class should be closer in the feature space. SVMs are efficient learning approaches for training classifiers based on several functions like polynomial functions, radial basis functions, neural networks etc. It is considered as a supervised learning approach that produces input-output mapping functions from a labeled training dataset. SVM has significant learning ability and hence is broadly applied in pattern recognition. Decision tree classifier organized a series of test questions and conditions in a tree structure. In the decision tree, the root and internal nodes contain attribute test conditions to separate records that have different characteristics.

The database consists of 69 CT images 36 images contain benign nodules and 33 contain malignant nodules. Training testing ratio considered is 70%-30%. A total of 49 images are used for training and 20 images are used for testing.

The comparative study was done between three classifier models to come out with a better accuracy. The Receiver Operating Characteristic (ROC) of the different classifier results is shown in Table.2.

Experimental results

The performance of the segmentation of lung nodule in CT images using vector quantization(VQ) is evaluated using the performance parameters along with their values are tabulated in Table 1.

Table 1. Performance of segmentation

Measures	Average values
Dice Co-efficient	0.66174
Jaccard Distance	0.9984
Mahalanobis Distance	190.2628
Hausdroff Distance	4.580666
sensitivity	0.99212
Area error rate	0.9974

A total of 69 lung CT images containing 36 benign images and 33 malignant images are used in this work. 10 images containing benign nodules and 10 images containing malignant nodules are used for testing. The accuracy, sensitivity, specificity and precision of the classification are calculated using the following equations.

$$Accuracy = \frac{\mathrm{TN+TP}}{\mathrm{TN+FN+FP+TP}} \tag{23}$$

$$sensitivity = \frac{\mathrm{TP}}{\mathrm{FN+TP}} \tag{24}$$

$$specificity = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \tag{25}$$

$$precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{26}$$

Linear, polynomial of different orders and RBF kernels are used in SVM for classification. The performance of the SVM classifier is as shown in Table 3. It is clear that the results of polynomial and RBF kernels are comparable and they are better than the linear kernel results.

Table 3. The classification results of the SVM kernel functions

	LINEAR	POLYNOMIAL -1	POLYNOMIAL-2	POLYNOMIAL-3	RBF
Accuracy	95.45%	96.5%	96.74%	95.86%	96.92%
Sensitivity	100%	100%	90.09%	92.8%	100%
Specificity	90.90%	94.1%	100%	100%	94.1%
Precision	90.90%	90.90%	100%	100%	90.90%

The results obtained from KNN, SVM (RBF kernel) and Decision tree are tabulated in the Table 4. All the classifiers are having the sensitivity of 100% i.e., correctly classifying all the malignant nodules as malignant.

Classifier performance is more than just a count of correct classification, the ROC curves of SVM classifier with linear, polynomial and RBF kernels are as shown in Fig.5. Polynomial and RBF are performing better compared to linear kernel.

Table 4. Performance comparison of KNN, SVM and DT

	KNN	SVM (RBF)	DECISION TREE
Accuracy	93.02%	96.92%	98.7%
Sensitivity	100%	100%	100%
Specificity	94.1%	94.1%	90.90%
Precision	90.90%	90.90%	90.90%

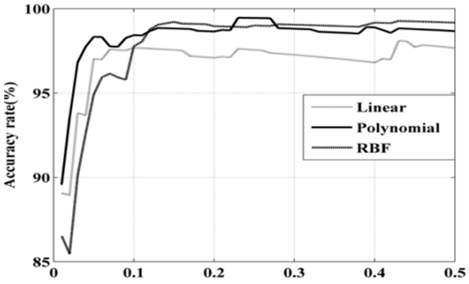

Fig.5 ROC curves of SVM

Comparison of the performance of KNN, SVM and DT in terms of ROC is as shown in Fig.6.

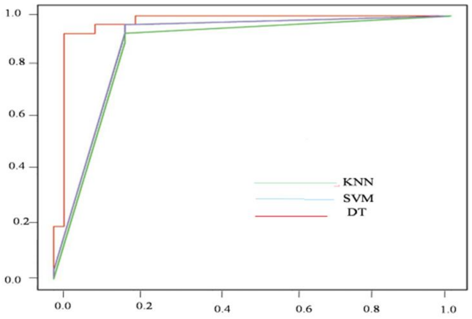

Figure.6 ROC curve of KNN, SVM and DT

Conclusion

This work presents a segmentation and classification method of lung nodules in CT images. Initially the preprocessing is carried out with an median filter in order to preserve and enhance useful information in the image. Then using Optimal thresholding and Vector quantization(VQ), the lung nodules are segmented efficiently. 11 textural features and 4 wavelet features are extracted for the classification purpose. KNN, SVM and Decision tree are used for classification of nodules. The sensitivity of all the classifiers is 100%. In SVM, polynomial and RBF kernels are better compared to linear kernel.

Acknowledgment

We would like to express our gratitude to Dr Rajesh, JSS Hospital, Mysuru, for his support in data collection and valuable suggestion to carry out this work.

References

[1] Chaudhary A & Singh, “Lung cancer detection on CT images by using image processing”. In Computing Sciences (ICCS), International Conference, pp. 142- 146, 2012.

[2] W. H. Organization, “Description of the global burden of NCDs, their risk factors and determinants,” Geneva, Switzerland:World Health Organization, ”, IEEE Transactions on Medical Imaging., vol. 9, pp. 7–19, 2011

[3] Sayani Nandy, Nikita Pandey “A Novel Approach of Cancerous Cells Detection from Lungs CT Scan Images’’ International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 8, August 2012

[4] S. Diciotti, G. et al., “3-D segmentation algorithm of small lung nodules in spiral CT images”, IEEE Transactions on Medical Imaging., vol. 12, no. 1, pp. 7–19,2008.

[5] Mr.Vijay A.Gajdhane, “Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques “, ISSN: 2278-8727, Volume 16, Issue 5, Ver. III, PP 28-3, Sep – Oct. 2014.

[6] Sunil Kumar et al., “Lung Segmentation using Region Growing Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering Volume 4. 2014.

[7] O. Talakoub, et al., “Lung segmentation in pulmonary ct images using wavelet transform,” in Acoustics, Speech and Signal Processing, IEEE International Conference on, vol. 1, pp. I–453, may-2012.

[8] Hao Han, Lihong Li, “Fast and Adaptive Detection of Pulmonary Nodules in Thoracic CT Images Using a Hierarchical Vector Quantization Scheme”, IEEE journal of biomedical and health informatics, vol. 19, no. 2,pp. 2168-2194, march-2015.

[9] E.Iyyapparaj, “A Study on Fast Adaptive Detection of Pulmonary Nodules in Thoracic CT Images Using a Hierarchical Vector Quantization Scheme”, International Journal of Innovative Research in Information Security, ISSN:2349-7017(O), Issue 06, Volume 3,September-2016

[10] Dr. H. B. Kekre, et al., “Detection and Demarcation of Tumor using Vector Quantization in MRI images”, International Journal of Engineering Science and Technology Vol.1(2), 59-66,April- 2009.

[11] Ada et al., ” Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images”. International journal of Advanced Research in Computer Science and Software Engineering, Vol. 3. 2013.

[12] Y. Shinagawa et al., “Stratified learning of local anatomical context for lung nodules in CT images,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2791–2798. 2010.