Middle-East Journal of Scientific Research 24 (Techniques and Algorithms in Emerging Technologies): 150-155, 2016
ISSN 1990-9233;
© IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.TAET23419

Statistical Feature-Based Support Vector Machine Approach for the Detection of Lung Cancer in Ct Images

[1]I. Jennetmary and [2]V. Ravi

[1]Electronics and Communication Engineering,
K.S.R. College of Engineering, Tiruchengode, India
[3]Electrical and Electronics Engineering,
K.S.R. College of Engineering, Tiruchengode, India

Abstract: Lung cancer is the growth of a tumor, referred to as a nodule that arises from cells lining the airways of the respiratory system. Nowadays cancer has become huge hazard in human life. There are many types of cancer, Lung cancer is one of the common types causing very high destruction rate. The best way of protect from lung cancer is its early detection and identification. The detection of lung cancer in early stage is a challenging problem, due to the structure of the cancer cells, where utmost of the cells are overlapping with each other. This computational procedure sorts images into similar clusters according to their similarities. In this Wiener filter is used for preprocessing of the images and GLCM feature extraction process and Support Vector Machine (SVM) classifier to check the condition of a patient in its different stages like stage-1, stage-2 or stage-3. The performance is based on the correct and incorrect classification of the classifier.

Key words: Respiratory system · Cancer cells · Support Vector Machine

INTRODUCTION

Lung cancer disease usually occurs if there is an unwanted growth in tissues of the lung. This growth spreads even beyond the lung and this process is known as metastasis. This spreads into other parts of the body too. Most cancers begin its growth in lung, named as carcinomas that are derived from epithelial cells. If an uncontrolled growth is detected at early stages, it helps to proceed with many treatment options, which reduces risk of invasive surgery and increased survival rate. The earlier the detection gives the higher the chances of successful treatment. It is considered to be the main cause of cancer death worldwide and it is hard to detect in its early stages because its symptoms appear only in the advanced stages causing the mortality rate to be the highest among all other types of cancer. Significant proof indicates that the early detection of lung cancer decreases the mortality rate [1, 2]. The latest estimates according to the recent survey provided by world health organization indicates that around 7.6 million deaths in whole world each year because of this type of cancer. Many techniques are available to diagnose lung cancer, like Chest Radiography (x-ray), Computed Tomography (CT), Magnetic Resonance Imaging (MRI) scan. These techniques are of high cost and time consuming one and detects the lung cancer at its advanced stages only. Hence, there is a great need of new technology to diagnose the lung cancer in its early stages. For manual analysis, Image processing techniques provide a good class tool. Artificial neural networks called as sixth generation of computing because it offers a complete problem solving approach. This reasearch aims at applying techniques like neural networks and their associated analysis to Health care, specifically to the management of lung cancer patients. A number of medical researchers utilized the analysis of sputum cells for early detection of lung cancer, most recent research relay on quantitative information, such as the shape, size and the ratio of the affected cells [3].

Lung Cancer: Lung cancer consists of a multiple growth of abnormal cells into a tumor. Cancer cells can be carried away from the lungs in blood or lymph fluid that

Corresponding Author: I. Jennetmary, Electronics and Communication Engineering,
K.S.R. College of Engineering, Tiruchengode, India.

surrounds by lung tissue [4, 5]. Metastasis refers to cancer spreading beyond its site of origin to other parts of the body.

Types of Lung Cancer: The initial stage of Cancer is called primary lung cancer. Two different types of groups named as Non- Small Cell Lung Cancer (NSCLC), Small Cell Lung Cancer (SCLC) are available.

Methodology: Lung cancer detection system (LCDS) system enhances the contrast and color of the images and then the nucleus present in the images are segmented with thresholding. Morphologic and colorimetric techniques are developed to extract features from the images of the nucleus. Neural network classifiers are employed for analyzing the features of the existence of cancer cells or not. Similarly, if there are cancer cells, the cancer cell type is identified.

Read input Image: Primarily, cancer and non-cancer patient's data or CT-Scan images [6] will be collected from different diagnostic centers. The digitized images are stored in the DICOM format with a resolution of 8 bits per plane.

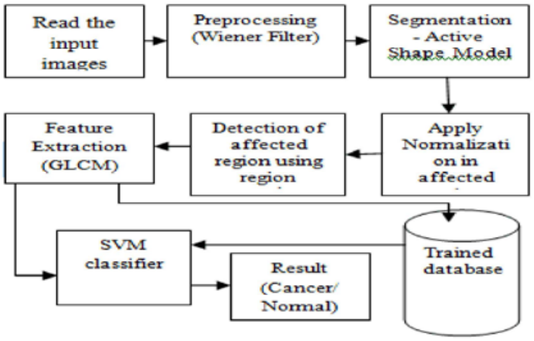

Read the input images
Preprocessing (Wiener Filter)
Segmentation - Active Shape Model
Apply Normalization in affected
Detection of affected region using region
Feature Extraction (GLCM)
Trained database
SVM classifier
Result (Cancer/ Normal)

Fig. 3.1: Methodology of work

Preprocessing: The image Pre-processing stage in this system begins with image enhancement which aims to improve the interpretability or sensitivity of information included in them to provide better input for other programmed image processing techniques. Image enhancement techniques can be divided into two wide types: Spatial domain methods and frequency domain methods. On the other hand, when image enhancement techniques are used as pre-processing tools for other image processing techniques, the quantifiable measures can determine which techniques are most suitable. In the image enhancement stage we will be using the wiener filter. The pre-processing of image aims for selective elimination of the redundancy in scanned images without affecting the original image, this play a vital role in the diagnosis of lung cancer. Therefore, wiener filter becomes the crucial step in preprocessing. Hence, each image is preprocessed to enhance its superiority.

Segmentation

Active Shape Model: The ASM model trains the images from manually drawn contours that find the main variations in the training data using Principal Component Analysis (PCA). This PCA enables the model with automatic recognition if it is a possible/good object contour. ASM modes consists of matrices that describe the texture of the lines perpendicular to the control point, these are used to correct the positions in the search step. After creating the ASM model, an initial contour is deformed by finding the best texture match for the control points. This is an iterative process, in which the movement of the control points is limited by what the ASM model recognizes from the training data as a "normal" object contour. Some possible ways to modify, refine and improve ASM are mentioned.

Bounds for the Shape Model: The shape model is fitted by projecting a shape in the -dimensional space (the number of landmarks, the factor two is because we consider 2-D images) upon the subspace spanned by the largest eigenvectors and by truncating the model parameters so that the point is inside the box bounded by. Thus there is no smooth transition; all shapes in the box are allowed, outside the box no shape is allowed. Clearly this can be refined in many ways, using a penalty term or an ellipsoid instead of a box and so on [1].

Nonlinear Shape Models: This type of shape model uses PCA and assumes the distribution of shapes in the - dimensional space as normal. If this is not true, non-linear (mixture models), could be more suitable (see for example).

Projecting the Shape Model: Based on (5), fitting the shape model, results in parameter minimization of the sum of squared distances between true positions and modeled positions. Practically it is desirable only to minimize the distance between true and model positions in a perpendicular direction to the object contour. This is because deviation along the contour does not change whether pixels are inside or outside the object. In it is demonstrated how to perform this projection on the contour.

Landmark Displacements: Behiels *et al*. proposed to use dynamic programming once the Mahalanobis distance has been computed at each new possible position to find new positions for the landmarks, instead of moving each point to the lowest distance position. This is to avoid the jumping of neighboring landmarks to new positions in different directions and thus leads to a “smoother” set of displacements. This can lead to quicker convergence.

Confidence in Landmark Displacements: If information is available about the confidence of the proposed landmark displacement, weighted fitting of the shape model can be used, as explained in. Initialization. Because of the multiresolution implementation, the initial position of the object (the mean shape, i.e., the mean location of each landmark) does not have to be very precise, as long as the distance between true and initial landmark positions is well within pixels. But if the object can be located anywhere within the input image, an (exhaustive) search to find a suitable initialization, e.g., as described in [2] can be necessary.

Optimization Algorithm: Hill climbing, Levenberg–Marquardt, or genetic algorithms are some standard non-linear optimization algorithms. This algorithm finds the optimal model parameters instead of using the alternating is placement algorithms of landmarks and model fitting. A minimization criterion could be the sum of the Mahalanobis distances, possibly complemented by a regularization term constructed from the shape model parameters. A multiresolution approach is used with standard non-linear optimization methods [5]. A shape model of a snake algorithm provides an internal energy term and the gray-level appearance model fit which is used as external energy term. This list is not complete, but it is beyond the scope of this article to present a complete discussion of the strengths and weaknesses of the ASM segmentation method [7]. The issues described above are not considered in this work. Instead, we focus on the following points:

Normalized First Derivative Profiles: This model is used to provide the original version of the gray-level appearance model. There is no a priori reason why this should be an optimal choice. In this paper, we propose an alternative.

Mahalanobis Distance: The Mahalanobis distance in (7) assumes a normal distribution of profiles. Practically, the distributions of profiles are not normal in cases in which the background of the object is one possible choice. The ASM scheme proposed to make use of a nonlinear classifier in the gray-level appearance model and can, therefore, deal with nonnormal distributions.

Training the Gray-level Appearance Model: i)Compute the 60 feature images for each training image. ii) At each resolution, for every landmark, a set of training samples are constructed with as input of 60 features and output as zero or one depending on whether the sample is in or outside the object. Samples are taken from an grid around the landmark in each training image. iii) For each training set, construct a NN classifier with selected optimal features. So the final result of the training phase is a set of classifiers.

ASM Procedure: i) Initialize with the mean shape ii)Start the coarsest resolution level. iii) For each landmark, put it at new locations, evaluate (9) with the NN classifier, move landmark to best new position. (iv) Fit the shape model to displaced landmarks. Iterate steps 3 and 4 times. If the current resolution is not yet the finest resolution, move to a finer resolution and go to Step 3.

Region Growing Method: The main goal of segmentation is to partition an image into regions. Some segmentation methods such as "Thresholding" achieve this goal by looking for the boundaries between regions based on discontinuities in gray levels or color properties. Region-based segmentation technique [8, 9] determines the region directly. Seed points consist of some basic concepts such as region growing which is to select a set of seed points. Seed point selection is based on some user criterion. The initial region begins with the exact location of these seeds. Then from this, the regions are grown to adjacent points based on a region membership criterion which can be an example of pixel intensity, gray level texture, or color. The image information is important and considers an example that if the criterion were a pixel intensity threshold value, knowledge of the histogram of the image would be of use, use, as one could use it to determine a suitable threshold value for the region membership criterion.

Feature Extraction: This is a special form of dimensionality reduction belongs to the pattern recognition and in image processing. If the input data is too large to be processed, it is suspected to be notoriously redundant and so the input data will be transformed into a reduced representation set of features (features vector). Feature Extraction is helpful in identifying brain tumor where is exactly located and helps in predicting next stage. Input data is transformed to the set of features called as feature extraction [2]. Some features are extracted by using GLCM [10] and Gabor are

Contrast: Contrast is defined as the separation between the darkest and brightest area.

$$Contrast = \sum_{i,j=0}^{n-1} P_{i,j}(i-j)^2$$

Correlation: Correlation is computed into what is known as the correlation coefficient, which ranges between -1 and +1.

$$Correlation = \sum_{i,j=0}^{n-1} P_{ij}\frac{(i-\mu)(j-\mu)}{\sigma^2}$$

Homogenity: Homogeneity is defined as the quality or state of being homogeneous.

$$Homogenity - \sum_{i,j=0}^{n-1} \frac{P_{ij}}{1+(i-j)^2}$$

Entropy: Entropy is a measure of the uncertainty in a random variable.

$$Entropy - \sum_{i,j=0}^{n-1} -In(P_{ij})P_{ij}$$

Energy: In GLCM, energy provides the sum of squared elements. Energy is the uniformity or the angular second moment.

$$Energy = \sum_{i,j=0}^{N-1} (P_{ij})^2$$

Shape: The term shape is commonly used to refer to the geometric properties of an object or its external boundary, as opposed to other properties such as color, texture, material composition.

Color: Color is a component of light which is separated when it is reflected off of an object. Colors are identified numerically by their coordinates.

Intensity: Intensity is a purity or strength of color.

Texture: It is the visual characteristic of a surface. For example, a surface can be rough or smooth.

Gray-level non Uniformity (GLN):

$$F_6 = GLN = \sum_{i=1}^{L}\left[\sum_{g=0}^{G-1} P(g,l)^2\right]$$

where l is the length of the run, L is the maximum run length, g is gray level bin, G is the maximum number of gray level bins and is the probability of the specific run length, respectively. This feature value increases as the gray-level outliers dominate the histogram.

Short Run Low Gray-level Emphasis (SRLGE):

$$F_7 = SRLGE = \sum_{G=0}^{G-1}\sum_{I=1}^{L}\frac{P(g,l)}{l^2(g+l)^2}$$

It is a diagonal metric that increases when the texture is dominated by many short runs of low gray value.

Long Run Low Gray-level Emphasis (LRLGE):

$$F_8 = \sum_{g=0}^{G-1}\sum_{I=1}^{L}\frac{P(g,l)l^2}{(g\,|\,l)^2}$$

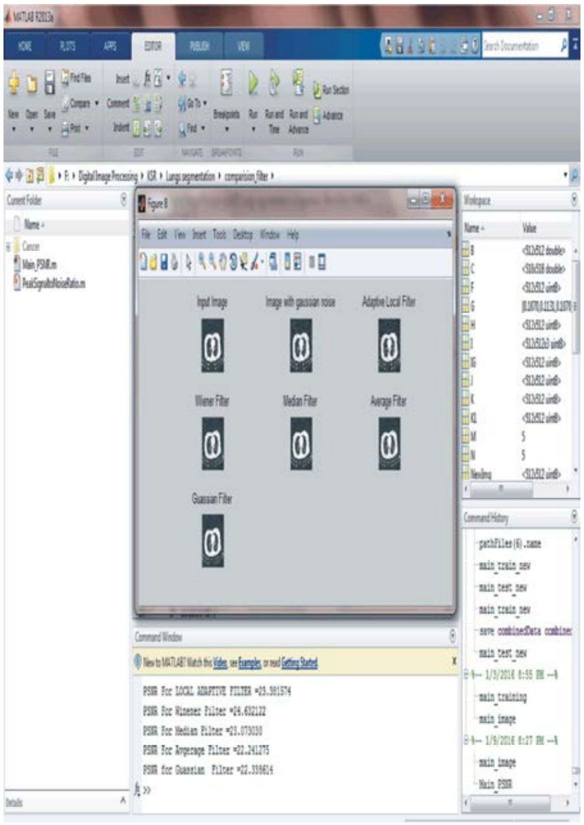
It increases when the texture is dominated by long runs that have low gray levels. Since each texture feature had different values in 4 scanning directions (Â=0Â°,90Â° and Â±45Â°), each texture feature computed in one CT image slice was represented by a mean of the 4 feature values calculated along 4 directions.

Support Vector Machine (Svm) Classifier: Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Properties: SVM belong to a family of generalized linear classifiers and can be interpreted as an extension of the perceptron. It can be considered a special case of Tikhonov regularization. A special property is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence they are also known as maximum. High accuracy, nice theoretical guarantees regarding overfitting and with an appropriate kernel they can work well even if you're data isn't linearly separable in the base feature space. Especially popular in text classification problems where very high-dimensional spaces are the norm.To go back to the particular question of logistic regression vs. decision trees (which I'll assume to be a question of logistic regression vs. random forests) and summarize a bit: both are fast and scalable, random forests tend to beat out logistic regression in terms of accuracy, but logistic regression can be updated online and gives you useful probabilities.

Experimental Results: Wiener filter is the proposed technique used for pre-processing. It is used to remove the noise in an image. It is better than mean filter, median filter, Gaussian filter [11]. In this segmentation technique, we are using an Active shape model (ASM) method and Region growing which gives more accurate result with help of the complement of segmented lung cancer image [12]. Using the ASM algorithm, it has an advantage of less computing time. In other words, the partitioned clustering is faster than the hierarchical clustering.Further it is also helpful for feature extraction. In this feature extraction technique we are use two different types of algorithm so it gives efficient result.

Preprocessing:



```
PSNR For LOCAL ADAPTIVE FILTER =23.381574
PSNR For Wiener Filter =24.632122
PSNR For Median Filter =23.073030
PSNR For Avgerage Filter =22.241275
PSNR for Guassian  Filter =22.338614
```

Fig: Preprocessing and PSNR values for different filters

The wiener filter is used for the preprocessing in the propossed work. Because it has a high PSNR value.

Segmentation:

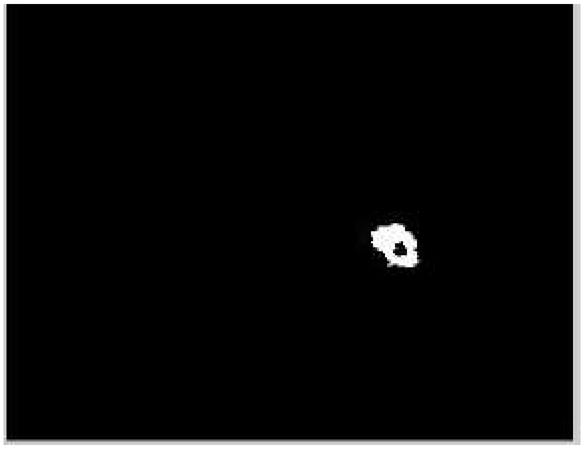
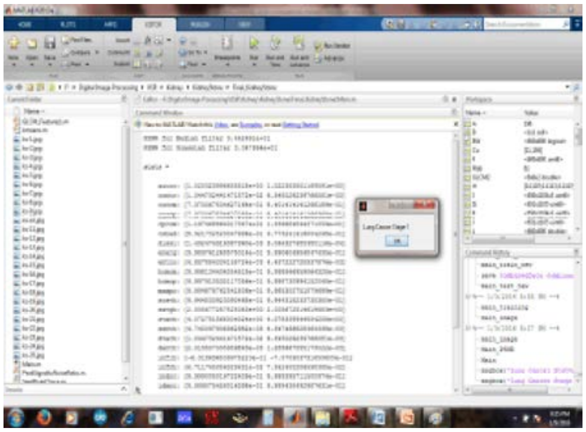


Fig: Segmentation using ASM and region growing

Features Extraction:



Lung Cancer Stage I

Fig: Features for segmentated lungs

Conclusion and Future Work: With the feature extracted, the proposed model identifies and detects the different stage of the disease. The initial process is to read the image and needs to preprocess because of high resolution and noise occur in the image where the noises are removed using median filter. The image is enhanced and segmented. Future work will be the preprocessing image will be the input for feature selection and extraction which are used to extract the particular region. The extracted features must be stored for classification. Based on the classification, stages will be identified which is used for a physician to give some therapy suggestions. The million order dataset can be selected and image classification can be done on larger dataset. With improved size of dataset various issues such as uploading data, managing feature set, increased execution time of classification algorithms etc. could be considered. More image features can be extracted for better classification Various combinations of previous features can be used to correctly classify medical data.

REFERENCES

1. Ada Rajneet Kaur, 2013. Early Detection and Prediction of Lung Cancer Survival using Neural Network classifier (IJAIEM), 2(6).
2. Vishukumar, S., K. Patela and Pavan Shrivastavab, 2012. Lung A Cancer Classification Using Image Processing", International Journal of Engineering and Innovative Technology, 2(3).
3. Jia, T., D.Z. Zhao and J.Z. Yang, 2007. Automated detection of pulmonary nodules in HRCT images, IEEE, 1st International Conference on Bioinformatics and Biomedical Engineering (ICBBE), pp: 833-836.
4. Dasu Vaman Ravi Prasad, 2013. Lung cancer detection using image processing techniques, International Journal of Latest Trends in Engineering and Technology
5. Sowmiya, T., M. Gopi, M. New Begin L. Thomas Robinson, XXXX. Optimization of Lung Cancer using Modern data mining techniques. International Journal of Engineering Research.
6. Sluimer, I.C., P.F. Van Waes and M.A. Vierever, 2003. Computer- aided diagnosis in high resolution CT of the lungs, Med. Phys., 30: 3081-3090.
7. Early Lung Cancer Action Program (ELCAP) pp: 234. 2014.
8. Diciotti, S., G. Picozzi and M. Falchini, 2008. 3D segmentation algorithm of small lung nodules in spiral CT images, IEEE Trans. Inf. Technol. Biomedical, 12: 7-19.
9. Sun, S.S., H. Li and X.R. Hou, 2007. Automatic segmentation of pulmonary nodules in CT images, IEEE, 1st International Conference on Bioinformatics and Biomedical Engineering (ICBBE), pp: 790-793.
10. Lung Database Consortium (LIDC): https:// imaging.nci.nih.gov/ ncia/login.jsf
11. Serge, V. Fotin, *et al*., 2009. A multiscale Laplacian of Gaussian filtering approach to automated pulmonary nodule detection from whole-lung low-dose CT scans, Medical Imaging 2009, Proc. of SPIE, 7260: 72601Q-1-8.
12. Chen, *et al*., 2004. Pulmonary micro nodule detection from 3D chest CT medical image, in MICCAI, 3217: 821-828.