International Journal of Engineering & Technology, 7 (2.21) (2018) 301-305

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



CAD system: a content based image retrieval approach for pulmonary nodule detection in CT images

Sajeev Ram[1*], Shylaja[2], Arun Sahayadhas[3]

[1]*Department of Computer Science & Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India.*
[2]*Department of Computer Science & Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India.*
[3]*Department of Computer Science & Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India.*
**Corresponding author E-mail:imsajeev@gmail.com*

Abstract

This paper proposes Computerized Aided Detection System (CAD) which uses Content Based Image Retrieval (CBIR) to detect cancer nodules present in an image. The CAD system is concerned for the radiologists to identify lung cancer at premature stages, which are very tiny nodules that are not able to seen by naked eye. In modern years, Image processing techniques play a key role in predicting diseases at early stages in particular in various cancer types such as liver cancer, breast cancer etc. This paper comprises of four steps: i) preprocessing an image in order to lessen the noise level and the accuracy of the image is to be improved, so that the accuracy in detection will be higher. ii) The image is segmented based on Marker-Controlled Watershed Segmentation. iii) The features of the nodules present in the image are extracted using GLCM. iv) The nodules are classified based on the extracted features using KNN classifier. The Content Based Image Retrieval Technique is used which is used to redeem query based images in the database by combining feature extraction and similarity matching methods. For experimentation of proposed technique, CT images are used which are extracted from Lung Image Database Consortium database (LIDC).

Keywords: *Marker-controlled watershed segmentation, KNN classifier, GLCM, CBIR, LIDC, sobel edge detection.*

1. Introduction

Lung Cancer is one of the serious causes of the cancer death among men and women. Every year, many people die due to lung cancer rather than other cancer like skin cancer, breast cancer, bladder cancer, bone cancer etc[1]. The risk factors for lung cancer are smoking, exposure to radon, exposure to asbestos, air pollution, arsenic in drinking water, family history of lung cancer etc. The symptoms of lung cancer do not appear until the disease is in advanced stage. Sometimes, lung cancer can be detected at early stage as a result of tests taken for other medical conditions like pneumonia, heart disease, lung disease[2].

Some of the manifestation of lung cancer are continuous coughing, chest pain, hoarseness, weight loss, breath shortness, wheezing, and tierd feeling. The lung cancer is detected using the following tests: Sputum cytology, Chest x-ray, Computed Tomography (CT scan), positron emission tomography (PET) scan [3]. The detection of lung cancer at early stages will help the cancer patients to survive for longer time. In order to predict the lung cancer at early stages, the lung cancer is to be detected accurately using Computer Aided Detection System (CAD) which helps the radiologists to predict the lung cancer nodules correctly. The radiologists reading are the current clinical standard in diagnosing the presence of cancer nodules[4]. The CAD system must find out the features such as margin, size, shape and give the brief description about the nodules present in the image.

The proposed system is based on Content Based Image Retrieval (CBIR) which helps to find out the similar feature of the given image from the large database. The major advantage of CBIR is the radiologists is able to provide the patients with a similar images from the database which is earlier diagnosed thereby providing them in their process of interpretation[5].

2. Literature review

Several researches has been proposed and implemented for the detection of lung cancer using various image processing techniques. Wook-Jin Choi , Tae-Sun Choi [6] proposed nodule detection system based on hierarchical block classification. For Image enhancement median filter and Gaussian filter are used. The SVM Classifier is used for extracting features. The system achieves 95.28% sensitivity. Ye, Xujiong Lin, Xinyu Dehmeshki, Jamshid Slabaugh, Greg Member, Senior Beddoe, Gareth [7] proposed a 3D local geometric and statistical intensity features for nodule detection. The detection rate of the sytem is 90.2% and false positive is 8.2/scan.

Taher, Fatma Werghi, Naoufel Al-ahmad, Hussain Sammouda, Rachid [2] proposed a system in which Hopfield Neural Network nad Fuzzy C-Mean algorithm are used for segmentation process. The threshold classifier is used to extract features. The accuracy of the system is 98%. The sensitivity of the system is 83% and the specificity of the system is 99%. Sharma, Disha Jindal, Gagandeep [8] proposed a system in which bitplane slicing is used as the extraction process and for segmentation region growing segmentation algorithm is used. The accuracy of the system is 80%.

Sharma, Disha Jindal, Gagandeep [9] proposed a system in which contrast enhancement, thresholding, filtering and blob analysis are used as preprocessing techniques. For segmenting an image, Otsu thresholding technique is used. Texture feature are extracted using artificial neural network(ANN). The accuracy of the system is

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

85%. Gattass, Marcelo Acatauassu, Rodolfo [10] proposed a system in which growing neural gas(GAG) technique is used to obtain the lung nodules present in an image. The support vector machine(SVM) is used to obtain the texture features of lung nodules. The sensitivity of the system is found to be 86%. The specificity is 91% and accuracy is 91%.

Lee, Y Hara, T Fujita, H Itoh, S Ishigaki, T [11] proposed a system in which genetic algorithm template matching(GATM) is used to detect lung nodules present in an image. The accuracy of the system is measured as 72%.Valdivieso, Manlio Amin, Hamdan [12] proposed a system in which algorithm template matching(GATM) is used to detect lung nodules present in an image. The features are extracted using lung nodule phantom images as reference image. The accuracy of the system is 90%.

Choi, Wookjin Cancer, Memorial Sloan-kettering Robust [13] used shape based feature descriptor to detect the lung nodules present in an image. The features are extracted and trained using SVM classifier. The sensitivity of the system is 97.5%. Kim, D Kim, J Noh, S Park, J Kim, J Noh, S Nodule, J Park Pulmonary[14] used gray level thresholding methods to segment an image. The features extracted using contour-follwing method. The sensitivity of the system is measured as 96%.

Kumar, Ashis Sudipta, Dhara Anirvan, Mukhopadhyay Garg, Mandeep Khandelwal, Niranjan [5] proposed a system in which content based image retrieval technique is used. The system works closer to SVM.

3. Proposed system

The proposed CAD system in which Content Based Image Retrieval (CBIR) is used to detect Pulmonary nodules in an CT image by analyzing the nodule database as well as feature database [5]. The first step is image acquisition. The second step is preprocessing an image so that noise present in the image can be reduced or removed. The third step is to segment an image, hence Marker-Controlled watershed segmentation method is applied to segment the image in order to identify the Region of Interest (ROI). The fourth step is feature extraction which measures certain properties or features and reduce the original dataset using GLCM feature extraction. The fifth step is feature selection based on genetic algorithm. The sixth step is classification which detects the lung nodules present in an image using KNN classifier and the content based image retrieval method is used. The Proposed architecture of the system is clearly shown in Fig.1.

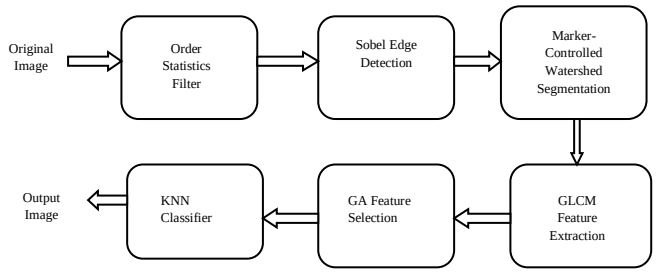


Fig. 1: Proposed architecture of CAD system

Image acquisition

The very first stage of any visual system is image acquisition. The input images are collected in the form of intensity, binary and RGB. Intensity represents brightness value. Indexed values are used as the index of a lookup table from which the value is read. RGB represents wavelength intensities. In the proposed CAD system LIDC images are used as input images. Fig.2 represents the collected image by acquisition process.

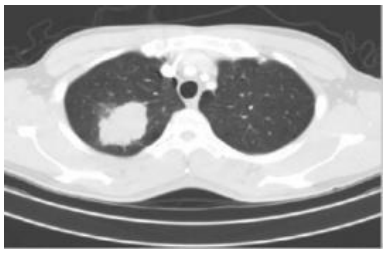


Fig. 2: Original image

Image pre-processing

Image preprocessing is a process of removing unwanted noise or distortions present in an image. In the CAD system, the image is converted into grayscale image. In the proposed system rgb2gray matlab code is used to convert original image into grayscale image[8]. This function converts the given input image into grayscale image by removing hue as well as saturation and retaining luminance. Fig.3 represents grayscale image.

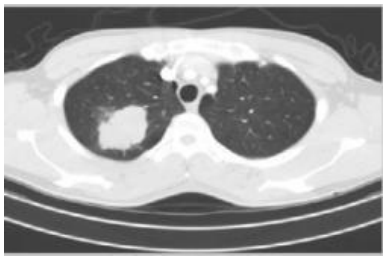


Fig. 3: Grayscale image

Order statistics filter

Order Statistics Filter is based on ordering pixels present in the image[15]. In the filtering algorithm, sliding window technique is used to perform pixel by pixel operation. The ordered statistical information is obtained by ordering neighbourhood data[16]. Some of the order statistics filters are median filter, max and min filter, midpoint filter, alpha-trimmed mean filter. Fig.4. represents Order Statistics Filter.

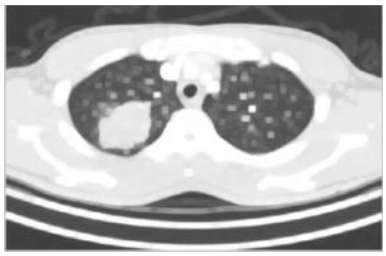


Fig. 4: Order statistics filter

Image segmentation

In the proposed system, Sobel edge detection algorithm is used to find the edges in an image [17]. To derive the edges of an image, edge detection method is used. The sobel method performs 2D spatial gradient measurement on images [16]. It is used to calculate gradient magnitude at each point in the input image. The sobel edge detection algorithm use 3x3 convolution masks. Fig.5 represent Sobel Edge Detection Algorithm

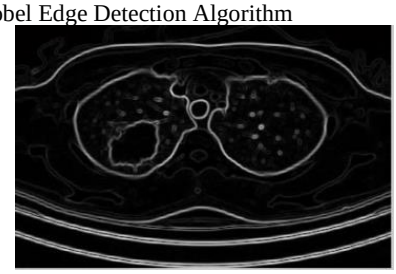


Fig. 5: Sobel edge detection algorithm

Marker-controlled watershed segmentation

In the proposed system, Marker-Controlled Watershed Segmentation is used to avoid calculation of many regions by reducing the minima [18]. The basis of this method are: i)morphological reconstructions. ii) Extracting the markers of the regions. iii) Applying watershed transform. This method is achieved using repeated morphological operations. In this method, the foreground and background marks are marked by opening and closing by reconstruction method. Due to this, we got limited number of region[19]. Fig.6. represents Marker Controlled Watershed Segmentation

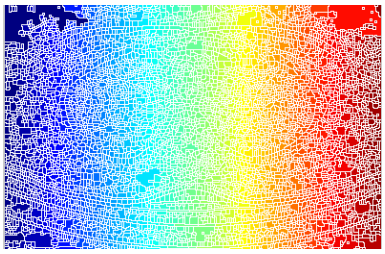


Fig. 6: Marker controlled watershed segmentation

Feature extraction

In feature extraction stage the original dataset is reduced by finding the features that distinguish one input image pattern from the other input image. The characteristics of the input type are provided to the classifier by considering the depiction of the significant properties of the image into feature space. In the proposed system, Grey Level Co-Occurance Matrix(GLCM) is used. The GLCM features is a second order statistical method to characterize texture spatial relationship in an image[20]. It is used to extract statistical parameter values such as energy, entropy, correlation, Inverse Direct Moment etc. In the GLCM matrix, the number of rows and columns will be equal to geey levels in an image[21]. The following statistical features are used for feature selection process:

- The contrast calculates the intensity of a pixel and its adjacent over the image. It is resolved by the divergence in the colour and brightness of the object and the further object within the equivalent field of view.

$$\text{Contrast} = \sum_{m=0}^{G-1}\sum_{n=0}^{G-1} Co(m,n)(m-n)^2$$

- The correlation is a measure of joint probability existence of the specified pixel pairs.

$$\text{Correlation} = \frac{\sum_{m=0}^{G-1}\sum_{n=0}^{G-1}(m-\mu m)(n-\mu n)P(m,n)}{\sigma_m \sigma_n}$$

- The cluster prominence measures the skewness or asymmetry in an image. When cluster prominence is less, there is an apex in the co-occurance matrix in the order of the mean values.

$$\text{Cluster Prominence} = \sum_{m=0}^{G-1}\sum_{n=0}^{G-1}\{m+n-\mu_i-\mu_j\}P(m,n)$$

- The cluster shade measures the skewness of an image and is defined by cluster shade feature. When cluster shade is high, the image is asymmetric.

$$\text{Cluster Shade} = \sum_{m,n=1}^{N}\left(m-\mu_i+n-\mu_j\right)^3 P(x,y)$$

- The dissimilarity measures the principled resemblance among segmentations produced by non-identical algorithms as well as segmentations on diverse images. The two feature vectors uses Texture features as its base. The relative frequency distribution is shown by feature vector and it is calculated by the Kullback-Leibler (K-L) divergence or relative entropy.

$$\text{Dissimilarity} = \sum_{a=1}^{I} f_{g,a} log \frac{f_{g,a}}{f_{q,a}}$$

- The energy is defined as the assess of the area of pixel pair repetitions. It calculates the consistency of an image using the formula:

$$\text{Energy} = \sum_{m=0}^{G-1}\sum_{n=0}^{G-1} Co(m,n)^2$$

- The entropy is used to exemplify the consistency of an input image. Its value will be utmost when every single one of the elements of the co-occurance matrix is same. Then entropy can be calculated using the formula:

$$\text{Entropy} = \sum_{m=0}^{G-1}\sum_{n=0}^{G-1} -Co(m,n)\log(Co(m,n))$$

- The Homogeneity calculates the proximity of the assessment of elements in the gray level matrix and it helps in enhanced segmentation. To acquire the local spatial statistics of the segmented image scaling is used and orientation characterizes the texture regions for similarity. The image is filtered out to a matrix of harmonized texture parts, and the texture features which are related with the regions are indexed in the image data.

$$\text{Homogeneity} = \sum_{m,n} \frac{P(m,n)}{1+|m-n|}$$

- The Sum of Squares, Variance put relatively high weight on the elements that vary from average value of P(i,j).

$$\text{Sum of Squares, Variance} = \sum_{m=0}^{G-1}\sum_{n=0}^{G-1}(i-\mu)^2 P(m,n)$$

```
The contrast of an image is 1.072400
The correlation of an image is 0.910006
The cluster prominence of an image is 610.658685
The cluster shade of an image is -3.168069
The dissimilarity of an image is 0.397655
The energy of an image is 0.227101
The entropy of an image is 2.129350
The homogeneity of an image is 0.871216
The maximum probability of an image is 0.355781
The sum of squares:variance of an image is 34.717284
The auto correlation of an image is 34.365542
```

Fig. 7: GLCM feature extraction

Feature selection

Feature selection is the process of removing irrelevant information while retaining significant information. Feature Selection plays a major role in pattern recognition problems such as image classification[22]. Feature Selection is used to reduce the dimensionality of feature space which will improve classification accuracy and time consumption can be reduced. In the proposed system, genetic algorithm(GA) is used for feature selection process[11]. The genetic algorithm operates in a binary search space and it operates on a sequence of 0's and 1's and it is called as chromosome. The proposed feature selection method is to find the feature set *S* which depends on the target class.

Image classification

The K- Nearest Neighbour (KNN) classifier is a classification technique used to find the distance between feature vectors so that we can predict and return an actual category of the image[23]. The training process for this classifier is storing feature vectors and labelling the training images. The unlabelled query is assigned to its k-nearest neighbours. In order to apply KNN classification, we need to know the value of the similarity function using a Euclidean distance formula:

$$Euclidean\ distance = \sqrt{\sum_{i-1}^{N}(qi-pi)^2}$$

Performance measures

In order to evaluate the proposed CAD system, some of the performance measures is to be taken place [22]. A true positive denotes abnormal lung nodule is accurately identified as abnormal. A true negative denotes normal lung nodule is accurately identified as normal. A false positive denotes normal lung nodule is erroneously identified as abnormal. A true negative denotes abnormal lung nodule is erroneously identified as normal. The confusion matrix [24] is used to calculate the performance of the CAD system. The table 1. Shows the confusion matrix of the

proposed system. It helps to visualize the performance of the learning system. Each row represents actual class and each column represents predicted class.

Table 1: Confusion matrix

Diagnosis	**Abnormal**	**Normal**
Abnormal	True Positive	False Positive
Normal	False Negative	True Negative

Sensitivity

Sensitivity correctly identifies the number of actual positives and measures its proportions. (i.e.) The percent of slice segmented comprising nodule is precisely identified as cancer nodule.

$$\text{Sensitivity}=\frac{True\ Positive}{True\ Positive+False\ Negative}$$

Here True Positive means slice segmented containing cancer nodules. False Negative means slice segmented containing nodule is precisely identified as non-cancer nodule.

Specificity

Sensitivity correctly identifies the number of negatives and measures its proportions. (i.e.) The percent of slice segmented without cancer nodule is precisely classified as non-cancer nodule.

$$\text{Specificity}=\frac{True\ Negative}{True\ Negative+False\ Positive}$$

Here True Negative means slice segmented without nodule is precisely identified as non-cancer.
False Positive means slice segmented without cancer nodule is precisely classified as cancer.

Accuracy

Accuracy identifies how classifier precisely works or excludes the situation.

$$\text{Accuracy}=\frac{\text{True Positive+True Negative}}{\text{True Positive+True Negative+False Positive+False Negative}}$$

Here True Positive means sliced segmented containing cancer nodules.
True Negative means slice segmented without nodule is identified as non-cancer.
False Positive means slice segmented without cancer nodule is identified as cancer.
False Negative means slice segmented containing cancer nodule is identified as non-cancer.
In this research, there are two types of train images are used. They are cancer image and non-cancer image. Totally we use 150 images to train the classifier. Of these 100 images are cancer image and 50 images are non-cancer image. The classification accuracy of the system is 98.2% and the sensitivity of the system is 97.1% and the specificity of the system is 96.5%.

4. Conclusion

In this paper we present a CAD system to detect the lung nodules present in LIDC images. The proposed system uses Marker Controlled Watershed Segmentation method to segment the image and the features are extracted using GLCM method. The feature selection process is carried out using Genetic algorithm. The content based image retrieval method is used to detect the cancer nodule present in the image by using KNN classifier technique. In the proposed system, hybrid methodology of combining genetic algorithm with KNN classifier is used so that the proposed method helps the radiologists in detecting lung cancer and take correct decisions. The proposed system has less positive values and the performance of the system is calculated using the statistical parameters. The system has 98.2% accuracy, 97.1% sensitivity and 96.5% specificity.

References

[1] Jazieh AR & Jazieh AR, "Lung Cancer Prevention and Early Detection".

[2] Taher F & Sammouda, R, "Lung cancer detection by using artificial neural network and fuzzy clustering methods", *IEEE GCC Conference and Exhibition (GCC)*, (2011), pp.295-298.

[3] Chen Y, Yang Z, Hu Y, Yang G, Zhu Y, Li Y, Chen W & Toumoulin C, "Thoracic low-dose CT image processing using an artifact suppressed large-scale nonlocal means", *Physics in Medicine & Biology*, Vol.57, No.9, (2012).

[4] Huang P, Park S, Yan R, Lee J, Chu LC, Lin CT, Hussien A, Rathmell J, Thomas B, Chen C & Hales R, "Added value of computer-aided ct image features for early lung cancer diagnosis with small pulmonary nodules: A matched case-control study", *Radiology*, Vol.286, No.1, (2017), pp.286-295.

[5] Kumar A, Sudipta D, Anirvan M, Garg M & Khandelwal N, "Content-Based Image Retrieval System for Pulmonary Nodules : Assisting Radiologists in Self-Learning and Diagnosis of Lung Cancer", *J. Digit. Imaging*, (2016).

[6] Choi WJ & Choi TS, "Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach", *Entropy*, Vol.15, No.2, (2013), pp.507-523.

[7] Ye X, Lin X, Dehmeshki J, Slabaugh G & Beddoe G, "Shape-based computer-aided detection of lung nodules in thoracic CT images", *IEEE Transactions on Biomedical Engineering*, Vol.56, No.7, (2009), pp.1810-1820.

[8] Sharma D & Jindal G, "Identifying lung cancer using image processing techniques", *International Conference on Computational Techniques and Artificial Intelligence*, (2011), pp. 115-120.

[9] Sharma D & Jindal G, "Computer Aided Diagnosis System for Detection of LungCancer in CT Scan Images", *International Journal of Computer and Electrical Engineering*, Vol.3, No.5, (2011).

[10] Netto SMB, Silva AC, Nunes RA & Gattass M, "Automatic segmentation of lung nodules with growing neural gas and support vector machine", *Computers in biology and medicine*, Vol.42, No.11, (2012), pp.1110-1121.

[11] Lee Y, Hara T, Fujita H, Itoh S & Ishigaki T, "Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique", *IEEE Trans. Med. Imaging*, Vol.20, No.7, (2001), pp.595–604.

[12] Dehmeshki J, Ye X, Lin X, Valdivieso M & Amin H, "Automated detection of lung nodules in CT images using shape-based genetic algorithm", *Computerized Medical Imaging and Graphics*, Vol.31, No.6, (2007), pp.408-417.

[13] Choi WJ & Choi TS, "Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor", *Computer methods and programs in biomedicine*, Vol.113, No.1,(2014), pp.37-54.

[14] Kim DY, Kim JH, Noh SM & Park JW, "Pulmonary nodule detection using chest CT images", *Acta Radiologica*, Vol.44, No.3, (2003), pp.252-257.

[15] Stella A & Trivedi B, "Implementation of Order Statistic Filters on Digital Image and OCT Image: A Comparative Study", *International Journal of Modern Engineering Research*, Vol.2, No.5, (2012), pp.3143-3145.

[16] Sharma N & Aggarwal LM, "Automated medical image segmentation techniques", *Journal of medical physics/Association of Medical Physicists of India*, Vol.35, No.1, (2010), pp.3-14.

[17] Sundaram M, Ramar K, Arumugam N & Prabin G, "Histogram-modified local contrast enhancement for mammogram images", *International Journal of Biomedical Engineering and Technology*, Vol.9, No.1, (2012), pp.60-71.

[18] Al-tarawneh MS, "Lung cancer detection using image processing techniques", *Leonardo Electronic Journal of Practices and Technologies*, Vol.11, No.21, (2012), pp.147-58.

[19] Kumar N & Arora S, "Medical Evaluation of Marker Controlled Watershed Technique for Detecting Brain Tumors", *International Journal of Computer Applications*, Vol.54, No.11, (2012).

[20] Depeursinge A, Van de Ville D, Platon A, Geissbuhler A, Poletti PA & Muller H, "Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames", *IEEE Transactions on Information Technology in Biomedicine*, Vol.16, No.4, (2012), pp.665-675.

[21] El-Baz A, Beache GM, Gimel'farb G, Suzuki K, Okada K, Elnakib A, Soliman A & Abdollahi B, "Computer-aided diagnosis systems

for lung cancer: challenges and methodologies", *International journal of biomedical imaging*, (2013).

[22] Parveen SS & Kavitha C, "A review on computer aided detection and diagnosis of lung cancer nodules", *International Journal of Computers & Technology*, Vol.3, No.3, (2012), pp.393–400.

[23] Bi L, Kim J, Wen L & Feng DD, "Automated and robust PERCIST-based thresholding framework for whole body PET-CT studies", *Annual International Conference of the Engineering in Medicine and Biology Society*, (2012), pp.5335-5338.

[24] Lawley MJ, Clarke BE, Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE & Colquist S, "Symbolic rule-based classification of lung cancer stages from free-text pathology reports", *Journal of the American Medical Informatics Association*, Vol.17, No.4, (2010), pp.440-445.