Classification of Benign and Malignant Tumors of Lung Using Bag of Features

A.Melody Suzan, G. Prathibha

Abstract— This paper presents a novel approach for feature extraction and classification of lung cancer, i.e., Benign or malignant. Classification of lung cancer is based on a code book generated by using Bag of features algorithm. In this paper 300 regions of Interest (ROI's) of lung cancer images from The Cancer Imaging Archive (TICA) sponsored by SPIE are used. In this approach Scale-Invariant Feature Transform (SIFT) is used for feature extraction and this coefficients are quantized using a bag of features into a predefined code book. This code book is given as input to the KNN classifier. The overall performance of the system in classifying tumors of lung is evaluated by using Receiver Operating Characteristics Curve (ROC). Area under the curve (AUC) is Az=0.95.

Index Terms— Bag-of-features, KNN, Lung Cancer, ROI, SIFT

1 INTRODUCTION

LUNG is the most common cancer in the world, leading to death. Abnormal cell multiplication of cell into the tumor is called lung cancer. Many treatment advancements like earlier detection through screening and increased awareness of lung cancer has reduced the death rate. American cancer society estimates lung cancer in 2016 are about 2, 24,390 new cases, i.e., 117,920 in men and 106,470 in women and 58,080 deaths from lung cancer [1].

There are different types of Lung cancers based on growing cell characteristics, broadly they are classified as:

1) Benign: Benign lung tumor is not cancerous, so will not spread to other parts of the body. They are normally smoother and more regularly shaped. It grows slowly and even stop growing or shrink. This type of tumors needs not to be removed and are not life-threatening. There are different kinds of benign lung tumors, the most common being hamatomas [2].

2) Malignant: Malignant lung tumor is cancerous and spreads to other parts of the body. They are often seen to have an irregular shape, rougher surface and color variation. About 85-90% of all lung cancers are malignant. There are three main types of malignant tumors adenocarcinoma, squamous cell carcinoma and large cell carcinoma which occur most often in smokers [3].

Research work has been done to classify the lung cancer type in earlier stage to improve the life span of the patient. Jing Z, Bin L, LianfangT, [4] proposed Lung Nodule Classification based on Wavelet and support vector machine which gives an accuracy of 84.89%. Zhang F.et.al [5] proposed Lung Nodule Classification with multilevel patch-based using SVM and probabilistic latent semantic analysis with an accuracy of 83%. Kumar S.A.et.al [6] proposed Lung Cancer Classification for CT images using Fuzzy Systems with an accuracy of 90%. Kuchunur used geometric features which are given as inputs to Artificial Neural Network (ANN) achieved an accuracy of 83% [7]. Kuruvilla and Gunavathi extracted statistical features and classified using Artificial Neural Network achieved an accuracy of 93% [8]. Chen H.et.al [9] proposed Lung Cancer Classification using Artificial Neural Network and Multivariable Logistic regression with an accuracy of 90%. Cascio [10] proposed CAD for selection of lung nodules in CT images with Detection rate of 88.5% and classification accuracy of 80%.

2 METHODOLOGY

There are different types of images used to diagnose the cancer, such as Chest Radiography (X-ray), Computed Tomography (CT), Magnetic Resonance imaging (MRI).The computerized cancer diagnosis comprises of three primary computational steps:

2.1 Preprocessing

Image preprocessing is a tool which uses many preprocessing operations like filtering, normalization, noise removal etc. It suppresses information that is not relevant to the specific image processing and enhance some image features important for further processing. The first step in image preprocessing is image cropping. Some irrelevant parts of the image can be removed by considering the image region of interest (ROI) [11].

2.2 Feature Extraction

The term feature can be stated as an "interesting" part of an image. Subsequent to image preprocessing, the features are extracted either at the tissue or cellular level. Quantification and distribution of the cells across the tissue are based on tissue level feature extraction. The properties of individual cells can be extracted based on cellular level feature extraction. There are many feature extraction techniques. In this paper Scale Invariant Feature Transform (SIFT) is used.

2.3 SIFT

It is an image descriptor developed by David Lowe (1999, 2004) for image based matching and Recognition. SIFT is invariant to translations, rotations and scaling transformations in the image and robust to illumination variations. It is used for detecting interesting points from gray-level image. These are obtained from Scale-Space Extrema of difference-of-Gaussians (DoG) within a difference -of-Gaussian Pyramid. By repeated smoothing and subsampling of an input image a Gaussian pyramid is constructed. And the difference-of-Gaussian is

computed from the difference between the adjacent levels in the Gaussian pyramid. The points at which the difference-of-Gaussian values assumes extrema with respect to both spatial coordinates in the image domain and the scale level in the pyramid are considered to be the points of interest. The difference between two filtered images is calculated known as the difference of Gaussian (DoG), D (x, y, σ), one with k multiplied by the scale of the other.

$$D(x,y,\sigma) = L(x,y,k\sigma) - L(x,y,\sigma) \quad (1)$$

Where $L(x,y,\sigma)$ is obtained by convolution of Gaussian Function $G(x,y,k\sigma)$ with an input image, $I(x,y)$.

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \quad (2)$$

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\exp\left\{-\left(x^2+y^2\right)\Big/2\sigma^2\right\} \quad (3)$$

By using DoG, the local maxima and minima of D(x, y, σ) can be detected by comparing each point with the pixels of all its neighbours.If this value is the minimum or maximum this point can be treated as an extrema. After calculating this points candidate list of Keypoints are found by eliminating some points that have low contrast or are poorly localized on an edge. The value of the Keypoint of the DoG pyramid at the extrema is given by:

$$D(Z) = D + \frac{1}{2}\frac{\partial D^{-1}}{\partial x}Z \quad (4)$$

If z value is below the threshold level the point will be excluded.

2.4 Bag of Features

By treating image features as a words Bag of features a can be applied to image classification. To represent an image as a bag of features it is treated as a document in which similar features are detected. For this, firstly features detection, Features description and code book generation are done. Features can be extracted from many techniques and best feature descriptor to represent the local patches into numerical vector is SIFT. Bag of features converts this numerical vectors into "code words". Code word can be generated by considering several similar features. By considering K-mean clustering code words are considered as centers of the learned clusters. Which produces a "code book". The code book size is the number of clusters. Each code word of the patch can be represented as histogram of the code word.

3 PROPOSED METHOD

3.1 DATASET

The dataset for this work is taken from SPIE with the support of AAPM (American Association of Physicists in Medicine) and the National Cancer Institute (NCI) [12].The dataset consists of lung cancer images of 403 malignant and 790 benign tumors in DICOM format for storage. In this Paper a subset of the database, i.e., 300 images which consists of 150 benign and 150 malignant are taken.

Fig. 1(a). Original images of benign lung cancer 0001

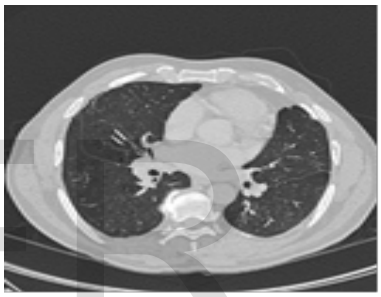

Fig. 1(b). Original image of malignant lung cancer 0082

These images are preprocessed to remove noise, normalization and cropped to obtain Region of Interest (ROI's) of size 256×256.This ROI's removes the unwanted pixel information and background. Figure 1 shows original images of benign and malignant lung images. Figure 2 shows ROI's extracted from different cases taken from the TICA-SPIE dataset.

Fig.2 (a). ROI of benign lung cancer 0001

- *A.Melody Suzan, pursuing master's degree in Electronics and Communication Engineering in Acharya Nagarjuna University, Guntur.*
- *G.Prathiba, Assistant Professor, Department of Electronics and Communication Engineering, Acharya Nagarjuna University, Guntur.*

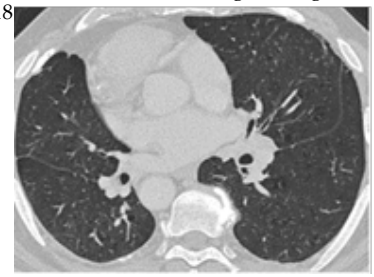

Fig.2 (b). ROI of malignant lung cancer 0082.

The proposed method classifies the lung cancer into two classes, namely benign and malignant. Before applying Bag of features, algorithm on this images, preprocessing is done to make this lung cancer images more suitable for extracting features. The ROI's are given as input to the SIFT descriptor.

In SIFT, First Scale Space Extrema is detected for series of difference of Gaussian (DoG) images obtained from original images at different scale for given ROI's of lung images.

From these keypoints are localized, the orientation of the keypoints is assigned and then rotation invariant descriptor using orientation was computed. These obtained features are given as input to the Bag-of-features, algorithm to build a dictionary (code book).

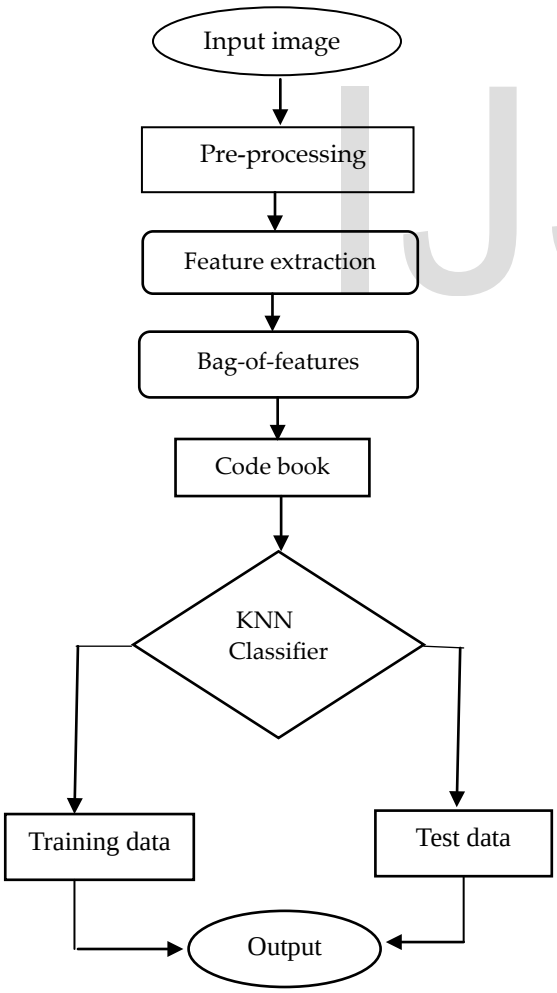

Fig. 3. Block diagram of proposed method.

Here to build code book, k-means clustering is done on this bag of features which contains huge data extracted from lung images.K-means clustering gives a centroid representation of feature group of each image. In this code book of size 500 and 250 is used

3.2 Classification:

There are different classifiers to classify the type of cancer based on features extracted. In this KNN classifier is used. It is identified as one of the top ten classifications by data mining method research community. It is a non-parametric for classification. Here the Euclidean distance between features vector of test images and feature vector of training image is calculated. Accuracy of the KNN algorithm depends on noise and unwanted features. Effort must be placed on selecting features. The code book generated using a bag of feature is given as input to the KNN classifier. The code book data are divided into 80% for training and 20% for testing. The Overall performance of the system is calculated using Receiver operating characteristics (ROC).

4 Experimental Results

The main objective of this method is to classify the lung cancer based on the given input images, subset of lung cancer images from SPIE.The ROI's of the lung images are given as input to SIFT and code book is generated using bag-of-features of size 250 and500.The code book is a histogram representation of lung images. This histogram image data is given as input to KNN classifier. The Accuracy of the system obtained is Az = 0.90% for code book of size 250 and Az=0.95% for code book of size 500.

n=300	Predicted BENIGN	Predicted MALIGNANT
Actual BENIGN	149	0
Actual MALIGNANT	14	137

Fig: 4 Confusion matrix for code book of size 250

n=300	Predicted BENIGN	Predicted MALIGNANT
Actual BENIGN	149	0
Actual MALIG-NANT	10	141

Fig. 5. Confusion matrix for code book of size 500.

The Confusion matrix for code book of size 250 and 400 are shown in Fig 4 and Fig 5.The Graph between True positive

rate and False positive rate is shown in Fig 6.The accuracy rate of code book of size 500 is more. The accuracy plot is shown in Fig 7.

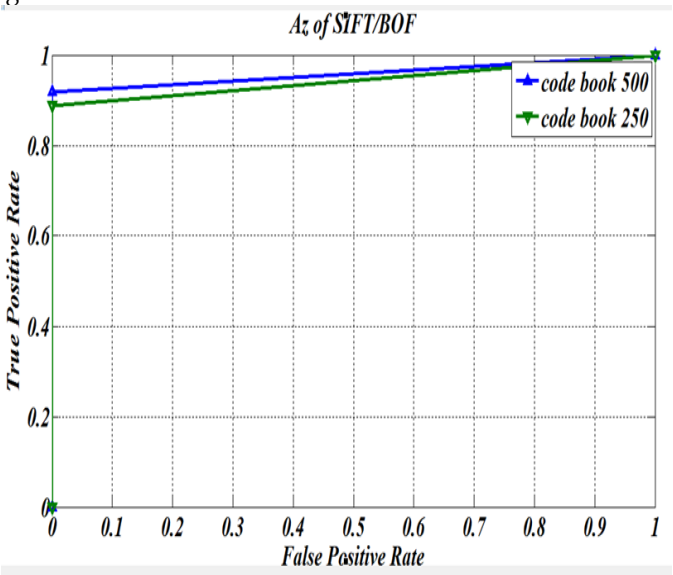

Fig 6 ROC curve for code book of size 250 and 500.

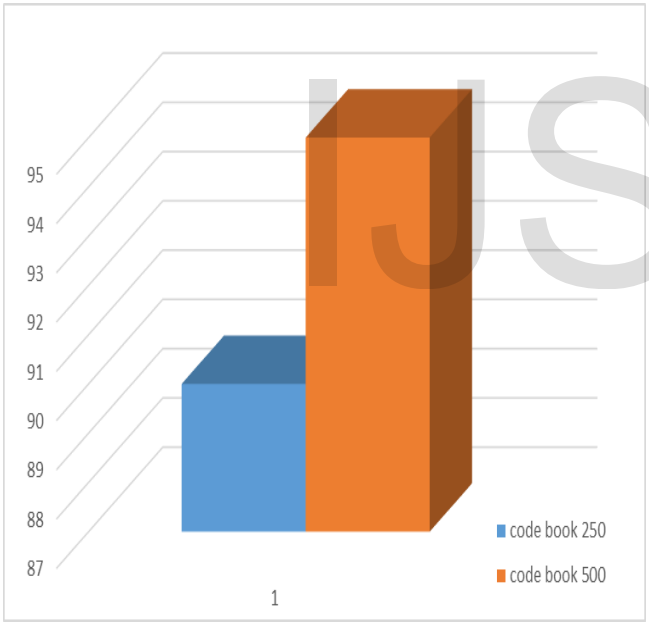

Fig 7 Accuracy rate for different code books.

5 Conclusion

The proposed method uses ROI's of Lung Cancer images for feature extraction using SIFT, for a further quantization Bag of features (BOF) is used to represent histopathological images obtained code of sizes 250 and 500 are given to a KNN classifier to classify the lung cancer. This method reports 95% accuracy for code book of size 500 and 90% accuracy for code book of size 250.Futher, we can improve the results by increasing code book size and considering only the cell area using various techniques.

ACKNOWLEDGMENT

Extending our grateful thanks to the authorities of Acharya Nagarjuna University College of Engineering and Technology for their support to utilize their facilities and encouragement to write this paper.

REFERENCES

[1] AmericanCancerSociety,https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html
[2] http://www.cancer.ca/en/cancer-information/cancer-type/lung/lung-cancer/non-small-cell-lung-cancer/?region=on
[3] http://my.clevelandclinic.org/health/articles/benign-lung-tumors
[4] Jing Z, Bin L, Lianfang T. Lung nodule classification combining rule-based and SVM. In: Edited by Li K, Proceedings of the IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications: 23–26 September 2010; Changsha, China. Piscataway, NJ: IEEE Computer Society; 2010. p. 1033–36.
[5] Zhang F, Song Y, Cai W, Lee M, Zhou Y, Huang H, et al. Lung nodule classification with multilevel patch-based Context analysis. IEEE T Bio-Med Eng. 2014; 61:1155–66.
[6] Kumar SA, Ramesh J, Vanathi PT, Gunavathi K. Robust and automated lung nodule diagnosis from CT images based on fuzzy systems. In: Edited by Manikandan V, Proceedings of the IEEE International Conference on Process Automation, Control and Computing: 20–22 July 2011; Coimbatore, India. Piscataway, NJ: IEEE Women in Engineering; 2011. p. 1–6
[7] S. A. Patil and M. B. Kuchanur, "Lung cancer classification using image processing," International Journal of Engineering and Innovative Technology, vol.2, no.3, 2012.
[8] Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for CT images," Computer Methods and Programs in Biomedicine, vol.113, no.1, pp.202–209, 2014.
[9] Chen H, Zhang J, Xu Y, Chen B, Zhang K. Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans. Expert Syst Appl. 2012; 39:11503–9.
[10] R.Bellotti, D.Casico, et al., ACAD System for nodule detection in low-dose lung CT's based on region growing and a new active contour modal, International Journal of Medical Physics and Practice 34(2007) 4901-4911.
[11] elib.mi.sanu.ac.rs/files/journals/kjm/32/kjom3209.pdf
[12] SPIE-AAPM-NCI Lung Nodule Classification Challenge Dataset, The Cancer Imaging Archive, 2 015.