

# LUNG NODULE CLASSIFICATION USING MULTILEVEL PATCH-BASED CONTEXT ANALYSIS AND DECISION TREE CLASSIFIER

M H Hasna<sup>1</sup>, Jobin Jose<sup>2</sup>

<sup>1</sup> M. Tech student (Applied electronics and instrumentation), Dept of ECE, Lourdes Matha College of Science & Technology, Trivandrum, kerala, India

<sup>2</sup> Assistant Professor, Dept of ECE, Lourdes Matha College of Science & Technology, Trivandrum, kerala, India

\*\*\*

**Abstract** - A novel classification method for the four types of lung nodules, i.e., well-circumscribed, vascularized, juxta-pleural, and pleural-tail, in low dose computed tomography scans is proposed. The method is capable to perform fully automatic segmentation and nodule detection from CT Scan Lungs images, based solely on information contained by the image itself. The proposed method is based on contextual analysis by combining the lung nodule and surrounding anatomical structures, and has three main stages. An adaptive patch-based division is used to construct concentric multilevel partition and feature set is designed to incorporate intensity, texture, and gradient information for image patch feature description. Finally, Decision tree classifier is utilized for classification of lung nodules. The system has been tested with a number of real Computed Tomography lung images and has achieved satisfactory results in classifying the lung diseases. The results show that the proposed method achieves higher classification performance than traditional methods.

**Key Words:** Lung nodule, Computed tomography, Decision tree classifier, Feature Design, and Patch-based division.

## 1. INTRODUCTION

Lung cancer is a major cause of cancer-related deaths in humans worldwide. Approximately 20% of cases with lung nodules represent lung cancers [1]. Therefore the identification of potentially malignant lung nodules is essential for the screening and diagnosis of lung cancer. Lung nodules are small masses in the human lung, and are usually spherical; however, they can be distorted by surrounding anatomical structures, such as vessels and the adjacent pleura [2]. Intraparenchymal lung nodules are more likely to be malignant than those connected with the surrounding structures, and thus lung nodules are divided into different types according to their relative

positions. Lung nodules can be divided nodules into four types: well-circumscribed (W) with the nodule located centrally in the lung without any connection to vasculature; vascularized (V) with the nodule located centrally in the lung but closely connected to neighboring vessels; juxta-pleural (J) with a large portion of the nodule connected to the pleural surface; and pleural-tail (P) with the nodule near the pleural surface connected by a thin tail. Sample images are shown in Fig 1, with the nodule encircled in red.

Computed tomography (CT) is the most accurate imaging modality to obtain anatomical information about lung nodules and the surrounding structures. In current clinical practice, however, interpretation of CT images is challenging for radiologists due to the large number of cases. This manual reading can be error-prone and the reader may miss nodules and thus a potential cancer. Computer-aided diagnosis (CAD) systems would be helpful for radiologists by offering initial screening or second opinions to classify lung nodules [3]. CADs provide depiction by automatically computing quantitative measures, and are capable of analyzing the large number of small nodules identified by CT scans.

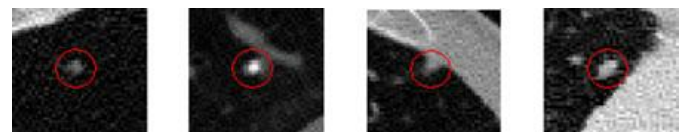


Fig -1: Transaxial CT images with the four types of nodules, shown from left to right, well-circumscribed, vascularized, juxta-pleural, and pleural-tail.

## 2. LITERATURE SURVEY

Fan Zhang *et al* [4] the paper titled as “Context Curves for Classification of Lung Nodule Images” In this paper the idea of classification of lung nodule is twofold. First, a superpixel labeling is conducted to label the context into foreground and background based on the patch division that groups related pixels together to represent certain anatomical structure. Second, a new feature descriptor, called context curve, is calculated according to the obtained labeling result by capturing the variation of

foreground pixels. the classification is conducted with SVM classifier based on the proposed feature descriptor.

Fan Zhang *et al* [5] the paper titled as “Lung Nodule Classification With Multilevel Patch-Based Context Analysis” In this paper classification of lung nodule is obtained by using support vector machine classifier. This method consist of three main stages: (1)adaptive patch-based division, it is used to construct concentric multilevel partition (2)new feature set is designed to incorporate intensity, texture, and gradient information for image patch feature description (3)SVM(support vector machine) based classifier is designed for the estimation of the relevant images

### 3. PROPOSED METHOD

This method consist of three main stages shown in fig 2: i) a patch-based image representation with multilevel concentric partition, ii) a feature set design for image patch description, and iii) a decision tree based classifier is used to estimate each lung nodule image. Concentric level partition of the image is designed in an adaptive manner with: (1) an improved superpixel clustering method based on quick shift is designed to generate the patch division; (2) multilevel partition of the derived patches is used to construct level-nodule (i.e., patches containing the nodules), and level-context (i.e., patches containing the contextual structures). Second, a feature set of three components is extracted for each patch of the image that are as follows: (1) a SIFT descriptor, depicting the overall intensity, texture, and gradient information; (2) a MR8+LBP descriptor, representing a richer texture feature incorporating MR8 filters before calculating LBP histograms; (3)a multiorientation HOG descriptor, describing the gradients and accommodating rotation variance in a multicoordinate system. Third, the category of the lung nodule image is finally determined by using decision tree classifier.

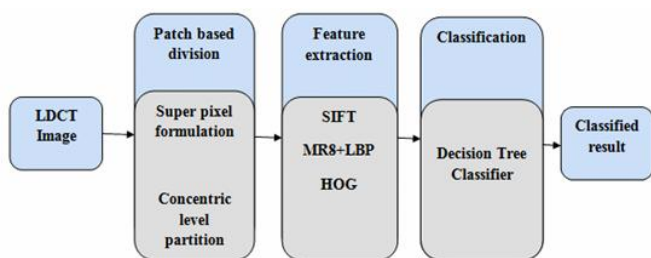


Fig -2: Block diagram representation of the proposed method

#### 3.1 Patch based division

Superpixel formulation is the process of dividing an image into multiple segments, which can incorporate local spatial information and reduce spurious labelling due to noise. In particular, quick shift is a mode seeking algorithm that can be used to partition an image into a set of superpixels

forming a tree of links to the nearest neighbour which increases an estimate of the density. However, due to the small size of lung nodules, a poor partition is often obtained when directly applying the quick shift method. To tackle this problem, employ quick shift in an iterative way with image amplification and down sampling. At the first stage, the image is amplified with nearest neighbour interpolation. A problem with direct use of quick shift on the original nodule image is that the image is so small that only a few pixels could present a particular anatomical structure, leading to the possibility of incorporating trans-regional pixels, as illustrated in Fig.3 (a). The bottom-left area in the sample image, corresponding to the white patch in (a), should be further divided because of the high contrast (white and grey). To obtain such an effect, amplify the original image based on the local intensity information. Quad-amplification generated the best performance by amplifying the image twice with twofold amplification each time through the experiments.

Then, the quick shift method is applied to the amplified image in an iterative way. Two parameters are introduced in quick shift[6]: *kernelsize*, the size of the kernel used to estimate the density, and *maxdist*, the maximum distance between points in the feature space that may be linked if the density is increased. Fixing them at particular values, i.e., the best performing parameter settings that obtain the highest classification rate with the standard quick shift generates too many patches, as shown in Fig.3(b). Dividing the image into too many patches would not only separate the integrated anatomical structure but also reduce the efficiency of the method, specifically during the feature extraction stage which extracts the feature set individually for each patch. Therefore, an iterative scheme is designed to handle this problem by increasing the initial values step by step to combine the clusters obtained in the previous iteration. Both *kernelsize* and *maxdist* were initialized at 2, increased by 0.3, and 1, respectively in each of the three iterations in the experiments. Finally, the down sampling stage is employed to restore the superpixel image to the original size. The clusters are thus the desired image patches. The illustration of the proposed superpixel formulation approach is shown in Fig 3.

Next, is to divide the patches in one image into multiple concentric levels, based on the distances between the patches and the centroid of the nodule patch. The nodule patch is the patch that contains the nodule centroid, which is given in the dataset. Level nodule is composed by the lung nodule patch, and level context is composed by

context patches, it represents different surrounding anatomical structures shown in fig 4.

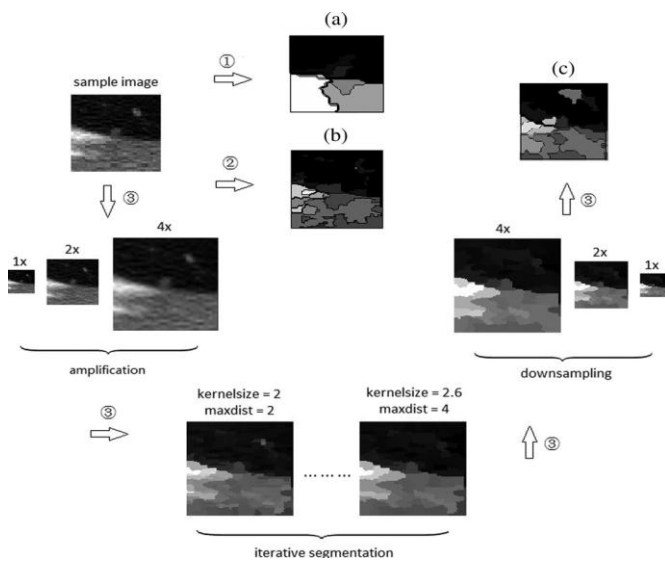


Fig -3: Procedure for the superpixel formulation scheme

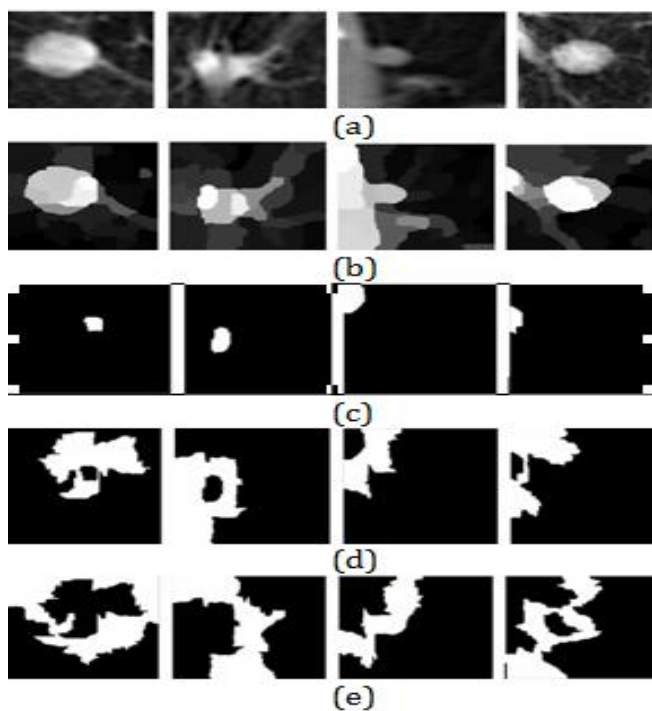


Fig -4: (a)Original image (b)Quick shift segmented image (c)level nodule (d)level context 1 (e)level context 2 of the well-circumscribed, vascularized, juxta-pleural, and pleural-tail

### 3.2 Feature extraction

Feature set is the combination of SIFT for overall description, MR8+LBP for texture, and multiorientation HOG for gradient. The SIFT (Scale invariant feature transform) process generates a 128-length vector for each key point. Since SIFT is invariant to image translation,

scaling, rotation and illumination changes, and robust to local geometric distortion, it provides valuable lung nodule data. Besides, it identifies the key points by computing extremum pixels in the image local area to incorporate the intensity information. Thus, SIFT descriptor was adopted as the first component of FS3 to give an overall description from intensity, texture, and gradient perspectives.

The combination of MR8 (maximum response 8) filters and LBP (Local binary pattern) [7] feature is designed to provide richer texture description of patches by incorporating multiscale and rotation-invariant properties. LBP is a powerful feature for texture based image classification. Although LBP can be easily configured to describe the local texture structure with multiresolution and rotation-invariance, it captures too many trivial image variations. Therefore, incorporate the MR filter set before computing LBP histogram. The MR set contains both isotropic and anisotropic filters at multiple orientations and multiple scales and records the angle of maximum response, which makes it possible to discriminate textures that appear to be very similar. MR8 bank consists of 38 filters but produces only eight filter responses by recording only the maximum filter response across all orientations for one scale. This yields rotation invariance. The final filter response contains two anisotropic filters for each of three different scales and two isotropic filters ( $2 \times 3 + 2$ ).

Gradient distribution provides helpful supplementary information to texture for discriminating various anatomical structures in nodule images. Among various gradient-based methods, HOG (Histogram of oriented feature gradient) is being widely used and can also improve performance considerably when coupled with LBP. However, unlike SIFT and MR8+LBP descriptors, the raw HOG descriptor cannot handle rotation-invariant problems. Therefore, designed a multiorientation HOG descriptor to provide further an advanced gradient description in addition to that from SIFT. The designed descriptor is adaptive to the locations of patches relative to the centroid of the nodule, rather than having the same initial orientation for all patches.

### 3.3 Classification

With the concentric level partition and feature set, the next stage is to label each image with one of the four nodule categories. The proposed method use a decision tree classifier for the classification of lung nodule. A decision tree is a decision support tool that uses a tree-like graph or model of decisions in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The path from root to leaf represents classification rules.

A decision tree consists of 3 types of nodes:

- Decision nodes - commonly represented by squares
- Chance nodes - represented by circles
- End nodes - represented by triangles

If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities. The decision tree can be linearized into decision rules, where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. Decision rules can also be generated by constructing association rules with the target variable on the right.

In computational complexity and communication complexity theories the decision tree model is the model of computation or communication in which an algorithm or communication process is considered to be basically a decision tree, i.e., a sequence of branching operations based on comparisons of some quantities, the comparisons being assigned the unit computational cost.

The branching operations are called "tests" or "queries". In this setting the algorithm in question may be viewed as a computation of a Boolean function  $f : \{0,1\}^n \rightarrow \{0,1\}$  where the input is a series of queries and the output is the final decision. Every query is dependent on previous queries. Several variants of decision tree models may be considered, depending on the complexity of the operations allowed in the computation of a single comparison and the way of branching. In this case use a liner decision tree for the classification.

Linear decision trees, just like the simple decision trees, make a branching decision based on a set of values as input. As opposed to binary decision trees, linear decision trees have three output branches. A linear function  $f(x_1, \dots, x_i)$  is being tested and branching decisions are made based on the sign of the function (negative, positive, or 0). Geometrically,  $f(x)$  defines a hyperplane in  $R^n$ . A set of input values reaching any particular nodes represents the intersection of the half-planes defined by the nodes.

#### 4. RESULT

The experiments were conducted on a number of real Computed Tomography lung images which contains 15 sets of low-dose CT lung image. Fig.5 shows the comparisons among the existing and proposed method. Existing method has three main stages. An adaptive patch-based division is used to construct concentric multilevel partition and feature set is designed to incorporate intensity, texture, and gradient information for image patch feature description. Finally, SVM classifier is utilized

for classification of lung nodules. But in proposed method use a decision tree classifier instead of support vector machine classifier. The existing method shows 60% accuracy for the classification of four types of lung nodule. But the proposed method using decision tree classifier shows 80% accuracy for the classification of lung nodule. The proposed method shows better result with 20% more of nodules correctly classified. Overall the proposed method shows best performance and it provide a promising ability for lung nodule classification.

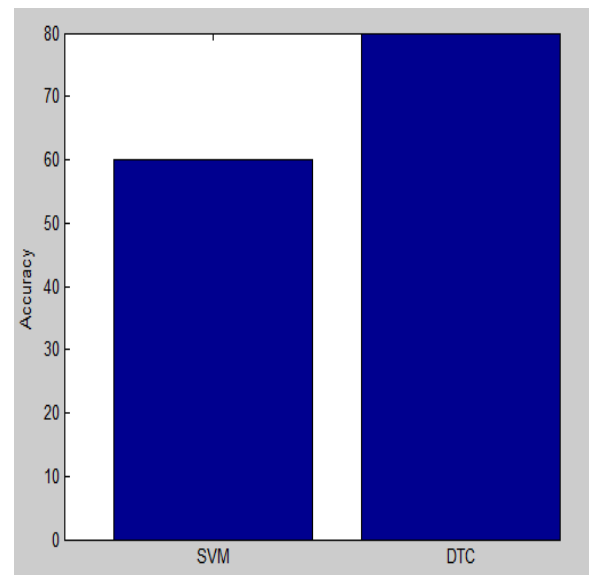


Fig -5: Comparison of accuracy between the SVM classifier and decision tree classifier

#### 5. CONCLUSIONS

This paper presents a supervised classification method for lung nodule LDCT images. The designed proposed method can overcome the problem of the lung nodule overlapping with the adjacent structures. The proposed system first preprocesses the image and extracts the features. These features were used for classification of lung nodules into four categories: juxta-pleural, well-circumscribed, vascularized and pleural-tail, based on the extracted information. Finally, Decision trees classifier are utilized for classification of lung nodules. The results from the experiments on the low-dose CT lung image are showed a promising performance for the proposed method.

#### REFERENCES

- [1] J. J. Erasmus, J. E. Connolly, H. P. McAdams, and V. L. Roggli, "Solitary pulmonary nodules: Part I. morphologic evaluation for differentiation of benign and malignant lesions," *Radiographics*, vol. 20, no. 1, pp. 43-58, 2000.
- [2] A. A. Farag, "A variational approach for small-size lung nodule segmentation," in *Proc. Int. Symp. Biomed. Imag.*, 2013, pp. 81-84.

- 
- [3] S. G. Armato III, M. L. Giger, and H. MacMahon, "Automated detection of lung nodules in CT scans: Preliminary results," *Med. Phys.*, vol. 28, no. 8, pp. 1552–1561, 2001.
- [4] F. Zhang, Y. Song, W. Cai, Z. Yun, S. Shan, and D. Feng, "Context curves for classification of lung nodule images," in *Proc. Int. Conf. Digital Image Comput.: Techn. Appl.*, 2013, pp. 185–191.
- [5] Fan Zhang, Yang Song, Weidong Cai, Min-Zhao Lee, "Lung nodule classification with multilevel patch-based context analysis," *IEEE Trans. Biomedical engineering*, vol. 61, pp. 1155–1165, April 2014.
- [6] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. ACM Multimedia*, 2010, pp. 1469–1472.
- [7] L. Srensen, S. B. Shaker and M. De Bruijne, "Quantitative analysis of pulmonary emphysema using local binary patterns," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 559–569, Feb. 2010.