

# Data-Driven Lung Nodule Models for Robust Nodule Detection in Chest CT

Amal A. Farag, James Graham, Salwa Elshazly and Aly Farag  
*Department of Electrical and Computer Engineering*  
*University of Louisville, Kentucky, USA*  
*E-mail: aafara02@louisville.edu*

## Abstract

*The quality of the lung nodule models determines the success of lung nodule detection. This paper describes aspects of our data-driven approach for modeling lung nodules using the texture and shape properties of real nodules to form an average model template per nodule type. The ELCAP low dose CT (LDCT) scans database is used to create the required statistics for the models based on modern computer vision techniques. These models suit various machine learning approaches for nodule detection including Bayesian methods, SVM and Neural Networks, and computations may be enhanced through genetic algorithms and Adaboost. The eminence of the new nodule models are studied with respect to parametric models showing significant improvements in both sensitivity and specificity.*

## 1. Introduction

Death due to lung cancer has greater rate than any other cancer type for both men and women. According to the Center for Disease Control (CDC) of the United States, in 2005 alone 90139 men and 69078 women in the United States died of Lung cancer. During that same year 89271 women and 107416 men were diagnosed with lung cancer. Lung cancer is the second most common cancer in the United States among white, African American and American Indian/Alaska Native men and women. Incidence rate per 100,000 people in the United States for men is 84.6 and 55.2 for women, while death rate for men is 69.4 and 40.6 for women [1]. The survival of lung cancer is strongly dependent of diagnosis [2][3]. Research studies to reach an optimal detection rate for early detection of lung cancer, is the hope for improved survival rate. Should the use of LDCT scans become a standard clinical practice an automatic way to analyze the scans will lend great benefit for the entire healthcare system; e.g., [4]-[6] and extensive survey in [7].

This paper describes an approach for generating non-parametric nodule models that capture the shape and texture information of the nodules. The ELCAP database [9] is used to test the approach. An image analysis approach for automatic detection and classification of lung nodules (e.g., Fig. 1) may be formed of four major steps: scan filtering to remove acquisition artifacts; segmentation to isolate the lung tissue from the rest of the chest region; nodule detection to isolate candidate nodules; and nodule classification which categorizes detected nodules into possible pathologies. Huge literature in place for the first three blocks (e.g., [7]). Due to space limitations, no extensive survey will be provided. This paper's main focus is on how nodule modeling will impact the detection and classification components in Figure 1. This paper is organized as follows: section 2 discusses the novel approach for template modeling and generation of the intelligent nodule templates; section 3 discusses performance evaluation; and section 4 concludes the paper.



**Fig. 1:** A Block diagram of the major steps involved in the computer-based analysis of LDCT of the chest in order to detect and classify doubtful lung nodules.

## 2. Nodule Simulation and Modeling

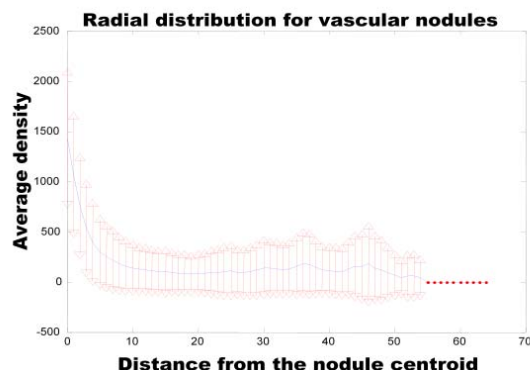
### 2.1 Pulmonary nodule definitions

In radiology a pulmonary nodule is a mass in the lung usually spherical in shape, however it can be distorted by surrounding anatomical structures such as the pleural surface. This paper uses the classification of Kostis et al [10], which groups nodules into four categories: vascularized where the nodule has significant connection(s) to the neighboring vessels while located centrally in the lung; well-circumscribed where the nodule is located centrally in the lung without being connected to vasculature; pleural tail where the nodule is near the pleural surface, connected

by a thin structure; and juxta-pleural where a significant portion of the nodule is connected to the pleural surface. These definitions will be adopted in this paper and the image analysis methods are developed and tested based on these nodule types.

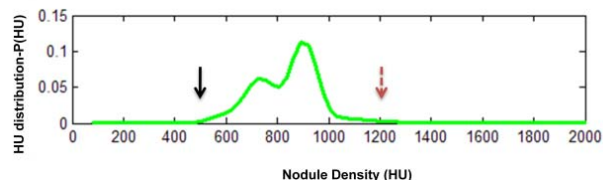
## 2.2 Nodule Simulation

The success for detection and classification is heavily reliant on proper nodule modeling. Modeling involves the shape, spatial support and the appearance (intensity) of the template. Various shapes and topologies can be taken by nodules in a CT scan, but the common characteristic amongst the nodules is the density distribution that tends to be concentrated around a region with an exponential decay (e.g., [6][9]). To illustrate this behavior, Fig. 2 shows the image intensity or Hounsfield Units (HU) vs. radial distance for the well-circumscribed nodule type in the ELCAP study. This distance was calculated by summing up the intensity values on concentric circles of various radii centered at the nodules centroid.



**Fig. 2:** Plot of the gray level density vs. radial distance from the centroid of the Vascularized nodule. The bars are one standard deviation off the mean values. Note the standard exponential behavior of the radial distance, this pattern has also been confirmed for the other nodule types.

Fig. 3 shows the average distribution of HU for the vascularized nodule types.



**Fig. 3:** Probability density of the Radial distance of the Vascularized nodule; Arrows show  $q_{min}$  and  $q_{max}$  of the range of densities.

Observations made for all nodule types in the ELCAP database used in this paper were similar to those made by Hu et al. (2001) [5] and Farag et al. (2006) [11] on different datasets, which is all of the nodule types

possessed the same characteristics of the radial, the HU or density decays exponentially with respect to the radial distance from the nodule's centroid. Furthermore, in the ELCAP dataset the decay of the HU is quite significant past a radial distance of 5 pixels. Hence, in designing a nodule template, we may specify a bounding box of size 10 pixels (corresponding to physical dimensions of 5mm, which is the range of interest for radiologists). In our experimentations we used templates of size 21x21 pixels. The information obtained about the nodule density distribution  $q_{min}$  and  $q_{max}$  for parametric templates, see figure 4, can be used to estimate the HU, at a distance  $r$  from the centroid using the following equations [9].

$$q(r) = q_{max} e^{-(r/\rho)^2}, 0 \leq r \leq R \quad (1)$$

$$\rho = R(\ln(q_{max}) - \ln(q_{min}))^{-0.5} \quad (2)$$

Where  $R$  is the radius of the circle, interior to the bounding box containing the nodule model (mean shape).

## 2.3 Statistical Nodule Modeling

The use of parametric templates has numerous drawbacks, such as the low sensitivity and unreliable specificity of the detected lung nodules. The new nodule modeling is based on analysis of shape and texture of candidate nodules selected by human experts. Developments in this paper uses 96 pre-identified nodules (24 nodules per type) by a bounding box of size 21x21 pixels (this region is based on the radial distance distribution). The ensemble of nodules contains variations in intensity distribution, shape/structural information and directional variability which the cropped regions within the determined bounding-box maintains. The 24 nodules per type are annotated to highlight the basic geometric and structural features of the nodules. We used Procrustes analysis to co-register those nodules with respect to any member of the ensemble, e.g., the first nodule can be used as the reference. Once each set of 24 nodules are co-registered the mean of the co-registered nodules is the template to be used in the subsequent analysis steps. Figure 5 visually describes the process undergone to generate the Vascularized nodule, and figure 6 shows the four non-parametric data-driven nodule templates designed.

As seen in Figure 6 the registration process leads to capturing the major features of each nodule type, thus the template is more descriptive of the particular nodule. This proves to be the key reason for the enhancements of the sensitivity and specificity of the

nodule detection process using template matching as will be shown in the next section.



**Fig. 4:** An ensemble of generated circular and semi-circular templates with various orientations.

Template matching as performed in our parametric template matching analysis [11] is now performed using the new data-driven templates. The behavior of the Normalized Cross-Correlation (NCC) for the new templates was studied by obtaining the NCC over all slices in the ELCAP study with known groundtruth for each nodule. The normalized cross-correlation of a template,  $t(x,y)$  with a sub-image  $f(x,y)$  is:

$$NCC = \frac{1}{n-1} \sum_{x,y} \frac{(f(x,y) - \bar{f})(t(x,y) - \bar{t})}{\sigma_f \sigma_t}, \quad (3)$$

where  $n$  is the number of pixels in template  $t(x,y)$  and sub-image  $f(x,y)$  which are normalized by subtracting their means and dividing by their standard deviations. The histogram for the average NCC for all nodule types is shown in Figure 7. The NCC behavior with the new nodule models takes the same general shape as with the parametric nodules [11] except the distribution function decays a lot faster as we approach a value of 0.5 – thus setting a threshold of 0.5 (to be able to compare with previous results) would result in detecting fewer nodules, but with better sensitivity and specificity as will be shown. In the implementation of the detection process, e.g., using template matching, we could use various orientations of the templates in Fig. 6.

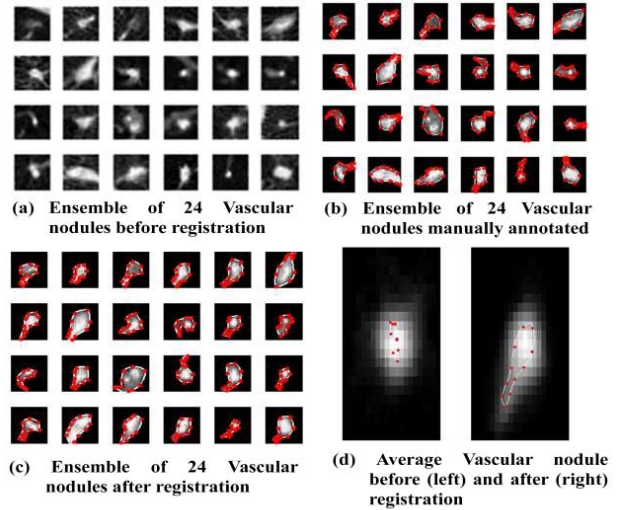
### 3. Performance evaluation

To test the effectiveness of the data-driven nodules with respect to the parametric models we implemented a decision fusion approach where *the four data-driven nodules* and the parametric templates, circular and semi-circular templates for various radius size and orientation, were used as the no-parametric and parametric templates in the detection stage, respectively. The templates for each template type (i.e. parametric and non-parametric) were used in a serial fashion and the final decision is the XOR of the four binary outputs. The output of the template matching from each nodule model is a binary image (NCC values rank from zero to 1; after thresholding the zeros are NCC values below 0.5 and the ones are otherwise). The tables below show some of the results obtained using the new intelligent nodules against the parametric templates. The candidate nodule detected is considered correctly detected and counted as true

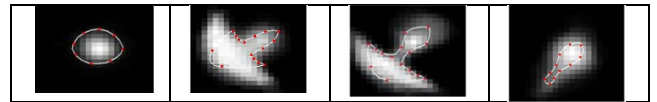
positive (TP) when the distance between the detected point and the closest ground truth point is smaller than the template radius. All other detected points are considered false positives (FP). The sensitivity and specificity were computed by the following equations:

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

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



**Fig. 5:** Generating the Vascular nodule starting from an ensemble of nodules.



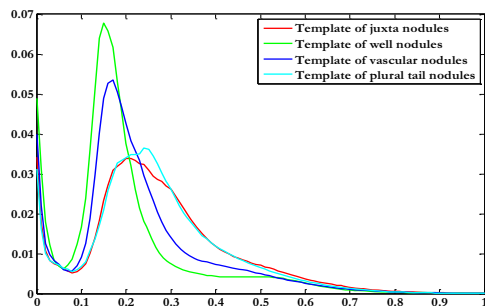
**Fig. 6:** The data-driven nodule models. From left to right: well-circumscribed, juxta-pleural, pleural tail and vascular nodule types. These models bear a great similarity to the true nodules.

### 4. Conclusions

In this paper, a data-driven approach was devised to model and simulate typical lung nodules. We studied the effect of template shape on detection of different nodules types. From our extensive experimentation we can conclude that the new data-driven models for template matching yielded an overall higher sensitivity and specificity rate than our previously used parametric templates. In the parametric case where we tested on all radii sizes between 1 and 20 pixels the sensitivity was higher but the specificity in comparison to the data driven nodule templates were still lower.

The overall performance depends on template shape and nodule type. The well-circumscribed nodule was the least sensitive nodule yet it emphasized the greatest improvement when the data-driven models

were used as shown in the above tables the sensitivity nearly doubled without increasing the specificity. The pleural tail in both the parametric and data-driven templates yielded the greatest sensitivity. Current efforts are directed to constructing and testing the new data-driven modeling approach on a large clinical data and extend this work into the 3D space.



**Fig. 7:** Distribution of the Normalized Cross-Correlation (NCC) for non- parametric templates. Higher NCC values results in less FPs while smaller values provide more FPs. Horizontal axis represents NCC value.

**Acknowledgements:** This work has been supported by the Kentucky Lung Cancer Research Program. A complimentary of this paper appeared in [12] and a detailed algorithmic evaluation is in the first author master’s thesis [13]. The first author has also been supported by a fellowship from the United States National Space and Aeronautics Agency, NASA.

## 5. References

[1] Centers for Disease Control and Prevention. <http://www.cdc.gov/cancer/dpcp/data/geographic.htm>

[2] A. Gajra, et al.: Impact of tumor size on survival in stage IA non-small cell lung cancer: a case for subdividing stage IA disease. *Lung Cancer* 42 pp.51--57 (2003)

[3] B. Zaho, G. Gamsu, M.S. Ginsberg, L. Jiang, L.H. Schwartz, “Automatic Detection of small lung nodules on CT utilizing a local density maximum algorithm,” *Journal of Applied Clinical Medical Physics* 4 (2003)

[4] Armato, S. G. 3rd, Giger, M. L., Moran C. J., Blackburn, J. T., Doi, K., MacMahon H.: Computerized detection of pulmonary nodules on CT scans. *Radio Graphics* 19 pp.1303--1311 (1999)

[5] S. Hu, E. A. Hoffman and J. M. Reinhardt, “Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images,” *IEEE Transactions on Medical Imaging*, Vol. 20, pp. 490–498, 2001.

[6] Y. Lee, T. Hara, H. Fujita, S. Itoh and T. Ishigaki, “Automated Detection of Pulmonary Nodules in Helical CT Images Based on an Improved Template-Matching Technique,” *IEEE Transactions on Medical Imaging*, Vol. 20, 2001.

[7] I. Sluimer, A. Schilham, M. Prokop, and B. van Ginneken, “Computer Analysis of Computed Tomography Scans of the Lung: A Survey,” *IEEE Transactions on Medical Imaging*, vol. 25, No. 4, pp. 385–405, April, 2006.

[8] ELCAP public lung image database, <http://www.via.cornell.edu/databases/lungdb.html>

[9] Amal A. Farag, S.Y. Elhabian, S.A. Elshazly and A.A. Farag, “Quantification of Nodule Detection in Chest CT: A Clinical Investigation Based on the ELCAP Study,” *Proc. of Second International Workshop on Pulmonary Image Processing in conjunction with MICCAI-09*, September 2009, pp. 149-160.

[10] W. J. Kostis, A.P. Reeves, D. F. Yankelevitz and C. I. Henschke, “Three dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images,” *Medical Imaging IEEE Transactions Vol. 22*, pp. 1259—1274, 2003.

[11] Aly A. Farag, A. El-Baz, G. L. Gimel'farb, R. Falk, M. Abou El-Ghar, T. Eldiasty, S. Elshazly, “Appearance Models for Robust Segmentation of Pulmonary Nodules in 3D LDCT Chest Images,” *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'06)*, Copenhagen, Denmark, October 1-6, 2006, pp. 662-670.

[12] Amal Farag, James Graham, Aly Farag and Robert Falk, "Lung Nodule Modeling – A Data-Driven Approach," 5th International Symposium on Visual Computing (ISVC09), Nov. 30 – Dec. 2, 2009, Las Vegas, Nevada, USA.

[13] Amal A. Farag, *Lung Nodule Modeling and Detection for Computerized Image Analysis of Low Dose CT Imaging of the Chest*, Master of Engineering Thesis, CVIP Lab, University of Louisville, April 2009.

**Table 1:** Performance of Template Matching for the data-driven vs. parametric templates for four nodule types, with improvements in sensitivity and specificity.

Nodule Type	Data-driven nodule models		Parametric Template with Radius = 10 and single orientation (0 °) for semi-circular models.	
	Sensitivity	Specificity	Sensitivity	Specificity
All nodule types	85.22%	86.28%	72.16	80.95%
Well-Circumscribed	69.66 %	87.10 %	49.44%	81.72%
Vascularized	80.4 %	87.0 %	70.73%	84.17%
Juxta-Pleural	94.78 %	86.54 %	83.48%	79.59%
Pleural-Tail	95.65 %	83.33 %	89.13%	79.33%

Nodule Type	Data-driven nodule models with orientation 0 ° -360 ° with step size 90 °		Parametric Template with Radius = 10 and single orientation 0 ° - 360 ° with step size 90 ° for semi-circular models.	
	Sensitivity	Specificity	Sensitivity	Specificity
All nodule types	80.07%	74.89%	78.01%	54.48%
Well-Circumscribed	62.92 %	76.20 %	51.69%	61.16%
Vascularized	65.85 %	75.53 %	75.61%	56.19%
Juxta-Pleural	93.04 %	73.16 %	92.17%	47.61%
Pleural-Tail	93.48 %	75.53 %	95.65%	57.14%