STATISTICAL MODELING OF THE LUNG NODULES
IN LOW DOSE COMPUTED TOMOGRAPHY SCANS OF THE CHEST

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

ABSTRACT

This work presents a novel approach in automatic detection of the lung nodules and is compared with respect to parametric nodule models in terms of sensitivity and specificity. A Statistical method is used for generating data driven models of the nodules appearing in low dose CT (LDCT) scans of the human chest. Four types of common lung nodules are analyzed using the Procrustes based AAM method to create descriptive lung nodules. Performance of the new nodule models on clinical datasets is significant over parametric nodule models in both sensitivity and specificity. The new nodule modeling approach is also applicable for automatic classification of nodules into pathologies given a descriptive database. This approach is a major step forward for early diagnosis of lung cancer.

Index Terms— Lung nodule modeling, Data-driven, Procrustes AAM Approach, Sensitivity and Specificity in CAD systems

1. INTRODUCTION

This paper deals with modeling of the lung nodules which appear in low dose computer tomography (LDCT) of the human chest. A number of screening studies in the US, Europe and Japan have been conducted in the past two decades for studying the enhancements of early detection of lung cancer using CT vs. X-ray and for studying the correlation of early detection and possible enhancement in lung cancer related mortality. Lung cancer is a worldwide problem. For example, in 2005 the Center for Disease Control (CDC) of the United States reported 90139 men and 69078 women death cases due to Lung cancer. During that same year 89271 women and 107416 men were diagnosed with lung cancer [1]. Lung cancer is the second most common cancer in the United States among white, African American and American Indian/Alaska Native men and women. The survival of lung cancer is strongly dependent on diagnosis [1]. 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 then an automatic way to analyze the scans will lend great benefit for the entire healthcare system.

An image analysis approach for automatic detection and classification of lung nodules may be formed of four major steps [5][8]: 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. The literature is rich in approaches to segment the lung from the rest of the chest tissues (e.g., [2]-[5]); but the majority of the nodule modeling methods are based on parametric descriptions of the nodules (e.g., in 2D circular or semicircular models are used, while in 3D volumes spherical or hemispherical models are used). Nodule detection is performed using various machine learning methods which execute template matching by one approach or another (e.g., [3][5]). Survey on automatic lung nodule detection may be found in [6].

This paper introduces a general approach for modeling lung nodules which is applicable for any nodule type and may be used in 2D slices or 3D volumes. The basic elements of the approach is as follows: 1) Generate an ensemble of lung nodules based on radiologists specifications; 2) Use computer vision methodologies to model the shape and texture information in the ensembles and generate a descriptive model for each nodule type; 3) Use the resulting nodule models for nodule detection and classification; and 4) Validate the effectiveness of the models in terms of sensitivity and specificity with respect to human experts.

Figure 1: The major steps in computer-based analysis of LDCT of the chest in order to detect and classify doubtful lung nodules [5][8].

To the best of our knowledge none of the worldwide lung screening studies has resulted in databases of nodules that identify their types and pathologies. Therefore, the need is persistent for reliable nodule models based on the actual scans; this is one of the goals of the authors of this paper, and a major effort is ongoing to achieve this goal. The availability of such database will lend great benefits for design and validation of proper nodule models that are “data-driven” and not hypothetical. A small clinical dataset
with labeled nodules from the ELCAP screening program [1] has been used to design and test the models in this paper, which show a great promise for the data-driven approach.

2. NODULE SIMULATION AND MODELING

2.1 Pulmonary nodule definitions

A pulmonary nodule usually appears in CT of the lung as a spherically shaped mass; however, this mass can be distorted by surrounding anatomical structures and the pleural surface. In this paper we use the classification of Kostis et al. [4], which groups nodules into four categories: juxta-pleural where a significant portion of the nodule is connected to the pleural surface; 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; and pleural tail where the nodule is near the pleural surface, connected by a thin structure; in all of these types there is no limitations on size or distribution in the lung tissue.

A database of nodules is constructed from the ELCAP lung screening project, using a semi-automatic method of cropping and categorizing each nodule into one of the four nodule types. Figure 2 shows an ensemble of each nodule type which will be used in the developments in this paper.

Figure 2: An ensemble of 24 nodules from the vascular (upper left), well-circumscribed (upper right), pleural-tail (lower left) and juxta-pleural (lower right) and nodule types.

2.2 Nodule Simulation

Modeling involves the spatial support, intensity (appearance) and shape of the nodule. There are various topologies and shapes that nodules may take as they appear in a CT scan. However, it has been experimentally observed that the density distribution of these nodules has an exponential decay behavior from the center (centroid) of the nodule (e.g., [3][5]). Figure 3 illustrates this behavior for the juxta-pleural nodule type in the ELCAP study, where we plot the image intensity (or Hounsfield Units) vs. radial distance.

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 qmin and qmax for parametric templates can be used to estimate the HU, at a distance r from the centroid using the following equations (e.g., [5]):

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

\[ \rho = R(\ln(q_{\text{max}}) - \ln(q_{\text{min}}))^{-0.5} \] (2)

where R is the radius of the circle, interior to the bounding box containing the nodule model (mean shape) and qmin and qmax determine the intensity range of a nodule in a given ensemble (e.g., Fig. 4).

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 nodule modeling approach presented here 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 21 x 21 pixels (this region is based on the radial distance distribution). The ensemble of nodules contains variations in intensity distribution, shape information and directional variability, which the cropped region within the determined bounding-box maintains.

(a)Ensemble of 24 Juxta nodules before registration
(b)Ensemble of 24 Juxta nodules manually annotated
(c)Ensemble of 24 Juxta nodules after registration

Figure 5: Generating the juxta-pleural nodule model.

The 24 nodules per type are annotated to highlight the basic geometric and structural features of the nodules. The mean nodule is generated per nodule type from the co-registered nodules. Various approaches may be employed for co-registration. In this paper we used the Procrustes approach we summarize below [7].

- **The Procrustes approach**

  Given two shapes \( x = (x_{11}, x_{21}), y = (\hat{x}_{11}, \hat{x}_{21}) \) \( i \in [0, N - 1] \). The Procrustes distance between \( x \) and \( y \) is a least-squares type metric of the form:

  \[ M^2 = ||x - y||^2 \]  

  \[ (3) \]

  Procrustes distance-based rigid registration between two shapes usually involves minimizing the expression \( ||T(x) - y||^2 \), where \( T \) in the Euclidean space is an affine transformation:

  \[ T(x) = [a - b \, x] + [t_x \, t_y] \]  

  \[ (4) \]

  The parameters \( a, b, t_x \) and \( t_y \) which provide minimum \( ||T(x) - y||^2 \) can be obtained by differentiation. We can show that these are:

  \[ a = \frac{x \cdot y}{||x||^2}, b = \left( \sum_{i} x_{2i} \hat{x}_{1i} - x_{1i} \hat{x}_{2i} \right) / ||x||^2 \]  

  \[ (5) \]

  \[ t_x = \left( \frac{1}{N} \right) \sum_{i} x_{2i}, \quad t_y = \left( \frac{1}{N} \right) \sum_{i} \hat{x}_{2i} \]  

  \[ (6) \]

- **The Active Appearance Model**

  The Active Appearance Model (e.g., [9]) is an approach to model shape and texture of an object using an ensemble. The approach is quite popular in computer vision and medical imaging analysis. This is the first time it has been used in lung nodule modeling. AAM falls under two forms: independent and combined AAM. In the case of independent AAM the shape and the appearance are separate parameters, thus, the modeling is performed separately, while combinational AAM consists of a single set of parameters which translates into simultaneous shape and appearance parameter calculation. The combined AAM case is used, where a combined set of parameters \( c = (c_1, c_2, \ldots c_i)^T \) parameterize the shape and appearance:

  \[ s = s_0 + \sum_{i=1}^{c_i} c_i s_i \quad \text{and} \quad A(x) = A_0(x) + \sum_{i=1}^{c_i} c_i A_i(x) \]

  The quantities in the above equations are estimated from the ensemble of pre-labeled nodules (i.e., Fig. 5 a). The annotated nodules are manipulated to extract the most discriminatory features for the shape \( s \) and appearance \( A(x) \).

  Figure 5 visually describes the process undergone to generate the juxta-pleural nodule. All nodules are co-registered with respect to the first element of the ensemble.

  Figure 6 shows the four nodule models that result from applying the Procrustes based AAM method as described above. These form the templates that we use for nodule detection by a template matching, for example.

  Figure 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.

  As shown in Figures 5 and 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.

### 2.4 Nodule Detection

To test the quality of the new nodule models, we performed nodule detection by template matching, where the templates are the show in Figure 6. Similar to other studies that used parametric templates (e.g., [3][5]) we also used the Normalized Cross-Correlation (NCC) as similarity measure. The normalized cross-correlation of a template, \( t(x,y) \) with a sub-image \( f(x,y) \) is:

\[ NCC = \frac{1}{n} \sum_{i,j} (f(x,y) - \bar{f})(t(x,y) - \bar{t}) \sigma_{f} \sigma_{t} \]

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 NCC behavior with the new nodule models takes the same general shape as with the parametric nodules [5] 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 Figure 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 were we used for the non-parametric templates the four data-driven nodules and for the parametric templates, circular and semi-circular templates for various radius size and orientation [5][8]. 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 table below show some of the results obtained using the new intelligent nodules against the parametric templates.

<table>
<thead>
<tr>
<th>Nodule Type</th>
<th>Data-driven models</th>
<th>Parametric Template with Radius = 10 and single orientation (0°) for semi-circular models.</th>
<th>Data-driven models with orientation 0°-360° with step size 90°</th>
<th>Parametric Template with Radius = 10 and single orientation 0°-360° with step size 90° for semi-circular models.</th>
</tr>
</thead>
<tbody>
<tr>
<td>All nodule types</td>
<td>Sensitivity</td>
<td>Specificity</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td></td>
<td>85.22%</td>
<td>86.28%</td>
<td>72.16</td>
<td>80.95%</td>
</tr>
<tr>
<td>Well-Circumscribed</td>
<td>69.66 %</td>
<td>87.10 %</td>
<td>49.44%</td>
<td>81.72%</td>
</tr>
<tr>
<td>Vascularized</td>
<td>80.4 %</td>
<td>87.0 %</td>
<td>70.73%</td>
<td>84.17%</td>
</tr>
<tr>
<td>Juxta-Pleural</td>
<td>94.78 %</td>
<td>86.54 %</td>
<td>83.48%</td>
<td>79.59%</td>
</tr>
<tr>
<td>Pleural-Tail</td>
<td>95.65 %</td>
<td>83.33 %</td>
<td>89.13%</td>
<td>79.33%</td>
</tr>
</tbody>
</table>

Table 1: Performance of Template Matching.

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 are defined in terms of the false positive (FP) and the true positive (TP) nodules.

4. CONCLUSIONS AND EXTENSIONS

In this paper, a data-driven approach was devised to model and simulate typical lung nodules. The new models yielded 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 was 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.

Acknowledgements: This research has been supported by the Kentucky Lung Cancer Program. The first author has been also supported by a fellowship from the United States National Space and Aeronautics Agency, NASA.

5. REFERENCES