Segmentation of nodules on chest computed tomography for growth assessment

William Mullally,[a] Margrit Betke, and Jingbin Wang
Computer Science Department, Boston University, Boston, Massachusetts 02215

Jane P. Ko
Department of Radiology, New York University Medical Center, New York, New York 10016

(Received 22 July 2003; revised 13 January 2004; accepted for publication 14 January 2004; published 16 March 2004)

Several segmentation methods to evaluate growth of small isolated pulmonary nodules on chest computed tomography (CT) are presented. The segmentation methods are based on adaptively thresholding attenuation levels and use measures of nodule shape. The segmentation methods were first tested on a realistic chest phantom to evaluate their performance with respect to specific nodule characteristics. The segmentation methods were also tested on sequential CT scans of patients. The methods' estimation of nodule growth were compared to the volume change calculated by a chest radiologist. The best method segmented nodules on average 43% smaller or larger than the actual nodule when errors were computed across all nodule variations on the phantom. Some methods achieved smaller errors when examined with respect to certain nodule properties. In particular, on the phantom individual methods segmented solid nodules to within 23% of their actual size and nodules with 60.7 mm^3 volumes to within 14%. On the clinical data, none of the methods examined showed a statistically significant difference in growth estimation from the radiologist. © *2004 American Association of Physicists in Medicine.* [DOI: 10.1118/1.1656593]

Key words: computer-aided diagnosis, lung cancer, shape analysis, 3D algorithms, phantom study, volume effects, computed tomography, image segmentation

I. INTRODUCTION

Lung cancer remains the leading cause of cancer death in the United States, with mortality of 160 000 people a year.[1] The overall 5-year survival rate is 15%,[2] but early detection and resection can improve patient prognosis. Low-dose CT is currently being evaluated as a possible screening method for the identification of early lung cancer.[3,4] Chest CT has been used to diagnose pulmonary metastases in oncology patients and evaluate the disease during treatment.[5]

A large number of patients undergoing screening for lung cancer have non-calcified nodules; approximately half of the nodules are *small*, usually defined to be less than 5 mm in diameter.[6] They are commonly benign but may also represent early malignancy.[6] Small nodules are therefore followed over time to determine potential size changes and evaluate growth rates.[7] Since the doubling time in nodule size is typically used as a measure of nodule malignancy in small nodules,[8] measurement accuracy is important for characterizing nodules as benign or malignant. To obtain accurate measurements of doubling times, methods for accurately segmenting nodules are needed.

The task of accurately segmenting nodules is challenging. To approximate nodule volume, radiologists typically report the size of a nodule in terms of its diameter in the axial plane, or in terms of its major and minor axes. The dimension of a nodule in the craniocaudal direction is generally overlooked. The partial volume effect makes it difficult for humans to classify voxels along tissue boundaries leading to large inter- and intra-observer variations in measurements,[9] especially in the volume calculation of small nodules. Table I demonstrates that over- or underestimating the diameter of a nodule by a single voxel can have large consequences in the volume calculation of the nodule. The errors compound when volume growth is estimated. A small object whose volume is overestimated in an initial scan and underestimated in a follow-up scan could actually be shown to shrink even though it doubled in volume.

Several techniques for nodule segmentation in the lung have been suggested.[10–13] They were applied after a radiologist had located the nodules and placed a region of interest around each. Zhao *et al.* examined two- and three-dimensional [(2D) and (3D)] nodule shape measures[11,12] and proposed the use of morphological filters to deal with vessels within the region of interest.[13] Ko *et al.* developed various methods to determine representative lung and nodule attenuation values.[10] A threshold halfway between these values was used to classify voxels in the region of interest as nodule or lung voxels.

Several papers have documented the use of phantoms, plastic materials that simulate how tissue is imaged on CT.[7,8,10,14] The true volumes of nodules in the phantom are known and can thus be compared to measurements obtained by visual inspection or automated methods.

Other works have focused on automatic nodule detection where nodules were delineated as part of the detection process.[15–22] One common approach to extract nodule candidate regions automatically was to impose multiple thresholds on the attenuation values within CT images.[16,19,21] Brown *et al.* used a threshold range to segment objects in the lung.[17] Gurcan *et al.* used a weighted *k*-means clustering algorithm

TABLE I. Comparison of volume to diameter of nodule measured on voxels of dimension 0.55 mm by 0.55 mm by 1.25 mm.

Diameter of object in voxels	Volume in mm^3	Diameter of object in voxels	Volume in mm^3
1	0.20	6	42.76
2	1.58	7	67.91
3	5.35	8	101.37
4	12.67	9	144.33
5	24.75	10	197.99

to separate nodules and vessels from background lung material.[20] Nodules were segmented slice by slice and assembled as 3D objects if the 2D components met several shape criteria.

In our work, we automated, extended, and compared several segmentation methods by Ko *et al.*[10] and Zhao *et al.*[11,12] In a preliminary study,[23] we extended the segmentation methods of Zhao *et al.* by selecting the region of interest automatically. In the current work, we further investigated this automated region-of-interest placement and applied it to 2D and 3D shape-based methods of Zhao *et al.* and to segmentation methods that extend the work of Ko *et al.* In particular, regions of interest were automatically created around all objects in the lung that were detected based upon their attenuation levels. The segmentation of these objects was then performed as if they were nodules even though many of them were composed partially or completely of other structures, for example, vessels. The strategy of our work was to separate the task of nodule segmentation from the task of nodule identification. Our methods segmented all lung objects and the task of identifying the nodules among them was then performed manually.

The focus of our work is automated analysis of nodule growth. There has been some preliminary work with clinical data by Ko and Betke,[21] Kawata *et al.*,[24] and Kostis *et al.*[25] The measurement of growth of small nodules is difficult to verify *in vivo* in humans, as the majority of these small nodules are not resected, and their true volumes and growth rates are therefore unknown. The use of a chest phantom provided us with an excellent scientific control for the case where the nodule growth rate was zero. For this important base case, we examined how the accuracy of automated growth estimation methods depended on nodule size and density. Phantom nodules of two densities, simulating solid and ground-glass nodules, were evaluated. Our tests on clinical data focused on solid nodules, which were mostly isolated. Some nodules were attached to vessels. We did not examine ground-glass or partially solid nodules.

The goal of our work is to develop a system that can be used for examining the growth rate of nodules in the clinical work flow. Such a system must be at least as reliable as any radiologist in measuring growth.

II. METHODS

Contiguous regions of soft tissue voxels in the lung are called ‘‘nodule candidates’’ in this paper. Nodule candidates are three-dimensional regions that can be segmented from the lung using a variety of techniques. Since the lung is composed of soft tissue and air, one way to identify nodule candidates is to impose some threshold, above which voxels can be considered to belong to a nodule candidate. The manner in which this threshold is selected affects the size and shape of nodule candidates. Additionally, because of the partial volume effect, voxels that fall above or below such thresholds cannot be assumed to correspond entirely to nodule or surrounding lung material.

The methods examined here used three different criteria for threshold selection and were classified as fixed threshold, variable threshold, and shape-based methods.

A. Fixed threshold segmentation

The simplest threshold method used a fixed threshold T_f throughout the entire lung. All voxels with a value greater than T_f were considered for nodule candidacy. Voxels were defined to belong to the same nodule candidate if they were 4-connected[26] to other voxel nodule candidates in the same axial slice. Voxels on different slices were considered connected if they were adjacent in the slices immediately above or below. The method tended to generate many small candidate regions that did not correspond to true nodules, so a filter was applied to remove small isolated groups of pixels from consideration.

The fixed threshold was used as a ‘‘seed value’’ that determined the initial nodule candidates for the variable threshold and shape-based methods detailed in the following. A range of seed thresholds was tested to discover how reliable the methods were given the choice of this seed value.

B. Variable threshold segmentation

Given the initial nodule candidates found by the fixed threshold method, a new threshold T_v was computed by finding representative nodule and lung tissue values and setting the new threshold halfway between these two values. Threshold T_v was then used to adjust which of the previously selected voxels belong to nodule candidates.

The variable threshold method presented here was based on a method by Ko *et al.*[10] They obtained representative nodule values by examining the slice image on which the nodule was most conspicuous and averaged the values of at least 20 voxels from the nodule's central region. For the smallest nodules where this was not possible, at least 5 voxels in the center were sampled. Five sample points were drawn from the lung and averaged to serve as the representative lung value. Two variants of extracting representative lung and nodule values were proposed by Ko *et al.* In the first variant, a representative nodule value was only drawn from the largest nodule found in the scan and the resulting threshold was applied to all nodules. A second variation extracted nodule values from each nodule so that each nodule's threshold was set independent of the other nodules in the scan.

It is important to note that the method by Ko *et al.* relied on hand-selected values. In our work, the approach was fully automated. Several variations in the automatic measurement of nodule and lung attenuation values were developed and tested. ‘‘Global values’’ were extracted based on the information in the entire lung region and applied to all nodules in the lung. ‘‘Local values’’ were extracted from each nodule candidate and its neighborhood and were only applied to each respective nodule candidate. The neighborhood of the nodule candidate was created by computing a minimum bounding box around the candidate and extending it by one voxel in all directions. Methods choosing average values and methods choosing the extreme minimum lung and maximum nodule candidate values were examined.

Two global lung values were automatically drawn from the entire lung by examining all non-nodule candidate voxels within the lung, i.e., voxels below the seed threshold. Lung value $l_{\mathrm{min},g}$ was the lowest non-nodule voxel value found in the lung; $l_{\mathrm{av},g}$ was the average of all non-nodule voxel values in the lung. Two global nodule values were also extracted from the largest nodule candidate within the lung. Nodule value $n_{\mathrm{max},g}$ was the largest voxel value within this largest nodule candidate; $n_{\mathrm{av},g}$ was the average of all voxel values in the nodule candidate core. Local nodule values $n_{\mathrm{max},l}$ and $n_{\mathrm{av},l}$ were drawn in the same manner from each nodule candidate individually. Similarly, two local measures for lung values were taken. Lung value $l_{\mathrm{min},l}$ was the lowest value found within the nodule candidate neighborhood. Lung value $l_{\mathrm{av},l}$ was the average of non-candidate voxels within the nodule candidate neighborhood.

Eight thresholds are derived from these representative values: $T_1=(l_{\mathrm{min},g}+n_{\mathrm{max},g})/2$, $T_2=(l_{\mathrm{min},g}+n_{\mathrm{av},g})/2$, $T_3=(l_{\mathrm{av},g}+n_{\mathrm{max},g})/2$, $T_4=(l_{\mathrm{av},g}+n_{\mathrm{av},g})/2$, $T_5=(l_{\mathrm{min},l}+n_{\mathrm{max},l})/2$, $T_6=(l_{\mathrm{min},l}+n_{\mathrm{av},l})/2$, $T_7=(l_{\mathrm{av},l}+n_{\mathrm{max},l})/2$, and $T_8=(l_{\mathrm{av},l}+n_{\mathrm{av},l})/2$. The notation $T_i^{(f)}$ is used in the following to indicate which seed threshold f was applied before representative lung and nodule values were extracted by method i.

C. Shape-based segmentation

Given an initial nodule candidate found by the fixed threshold method, a new threshold T_{sh} was computed on the basis of shape information. This threshold was then used to delineate the nodule.

Zhao *et al.*[11,12] used ‘‘gradient strength’’ and compactness criteria to refine the segmentation from an initial bounding box manually set around a nodule. The measures were automatically observed over a range of thresholds and served as a basis for choosing an optimal threshold with which to segment a nodule candidate. We re-implemented the two-dimensional approach of Zhao *et al.*[11] substituting their manually cropped regions of interest with regions automatically segmented with a seed threshold. We also implemented a variant of their 3D algorithm, using two measures, differences in 3D density values along the nodule border and sphere occupancy, to refine the segmentation we obtained from applying the seed threshold.

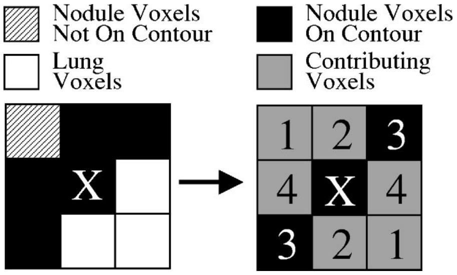

FIG. 1. Contour strength: Example of a voxel X on the nodule border illustrated in two dimensions. Only corresponding pairs 1, 2, and 4 contribute to the contour strength measure.

1. 3D Density difference measure

In this work, the two-dimensional measure of ‘‘gradient strength’’ proposed by Zhao *et al.*[11] was generalized to three dimensions and called ‘‘contour strength.’’ It was defined as the average magnitude of density differences across the nodule contour for every voxel on the nodule candidate’s contour. The contour of a nodule candidate was defined to contain all voxels with at least one neighbor that was not part of the nodule candidate. This neighborhood was defined by six connections in three dimensions.

At each voxel V_i under consideration, if a voxel V_j on one side of V_i was part of the nodule candidate and the voxel V_k on the opposite side of V_i was not part of the nodule candidate, then the contour strength of V_i was defined as the average of the absolute values of density differences between all such pairs of opposing voxels V_j and V_k. There are thirteen such possible pairs around each voxel V_i (see Fig. 1).

2. Sphere occupancy measure

The radius r of a nodule candidate was defined to be the distance from the candidate’s centroid to its furthest point. This radius was then used to define a sphere that completely encompasses the nodule candidate. By comparing the volume of an object with the volume of the sphere encompassing it, the portion of the sphere the object occupied was computed. In particular, given the volume V of the nodule candidate, the sphere occupancy of the candidate was defined as $c=3V/(4\pi r^3)$, where $0<c\leqslant 1$ (see Fig. 2). A nodule with spherical shape has an occupancy value of one. We set $c_{\mathrm{min}}=0.25$ as a desirable lower bound on the occupancy value of a nodule. This value represents a strong require-

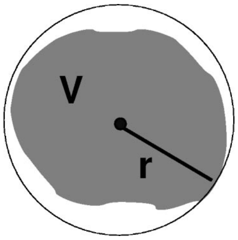

FIG. 2. Sphere occupancy: The radius r of an object can define a sphere to contain it. The ratio of this sphere to the volume V of the object is the sphere occupancy of that object.

ment. Consider, for example, adding a one-voxel protrusion to a perfectly spherical object of diameter 4 voxels and volume 12.67 mm^3 (see Table I). This would increase the object's diameter by 1 voxel, and the sphere occupancy of the encompassing sphere would decrease from 1 to 0.296. On the other hand, vessel structures were often found to have sphere occupancies close to 0.

3. Combining 3D measures for determining final shape-based segmentation

To find threshold T_{sh} both the contour strength and the sphere occupancy of the nodule candidate were computed over a range of n threshold levels $T_{s1},\ldots,T_{sn}$, beginning at the lowest value T_{s1}. In some cases a single object segmented with a low threshold broke into several objects at a higher threshold. When this happened, the shape-based methods operated on each object separately to produce a segmentation for each such object. At the low threshold where the objects were not distinct, the segmentations considered were the same for both objects. The segmentations considered for each object were only different at higher thresholds where the objects were distinct (see Fig. 6). In these cases, the shape-based method returned a segmentation for each object found in the candidate region based on the contour strengths and sphere occupancies of each object considered individually as described in the following. Because such objects were treated separately, the best segmentations of such objects could be at different thresholds.

For each object, the threshold T_{sk} that contained the maximal contour strength was considered first. If the object had a sphere occupancy value greater than $c_{\min}$ at T_{sk}, then T_{sh} was assigned to be T_{sk}. Otherwise a sequence of increasingly higher thresholds was checked until a threshold was found at which the object had a sphere occupancy value greater than $c_{\min}$. If no threshold satisfied this constraint, the threshold at which the highest sphere occupancy was found was considered to be T_{sh}. When multiple objects were found in a region of interest, if the best segmentation of each object was chosen at the threshold after which objects have become distinct, then these objects would not overlap. However, if the best segmentation of one or more of these objects was chosen at the threshold where objects were not yet distinct, then such segmentations would completely engulf the objects segmented at an equal or higher threshold. Where overlapping segmentations resulted, spurious segmentations were detected easily by visual inspection and removed.

D. Error measures

To measure which of m methods performed best in estimating nodule growth, the least-squares error between the growth $G_{i,j,k}$ of a nodule i measured by a method j in the CT scan reconstruction k and its true growth G_i^{true} was computed:

$$\min_{j\in\{1,\ldots,m\}} \sum_{k=1}^{s}\sum_{i=1}^{n}(G_{i,j,k}-G_i^{\text{true}})^2, \tag{1}$$

where s is the number of CT scan reconstructions and n the number of nodules. The growth of a nodule i was defined by comparing its volume $V_{1,i,j,k}$ in scan 1 with its volume $V_{2,i,j,k}$ in scan 2:

$$G_{i,j,k}=\frac{V_{2,i,j,k}-V_{1,i,j,k}}{V_{1,i,j,k}}. \tag{2}$$

The growth was not expressed as an absolute volume difference (as, for example, in Ref. 10), but instead as a relative measure, the ratio of the volume difference and the initial volume. This simplified the comparison of volume changes for nodules of different sizes and gave the same importance to small nodules as to large nodules in the error analysis.

1. Error measure on phantom data

Since the true growth G_i^{true} of the phantom nodules was zero, the least-squares measure in Eq. (1) reduces to $\min_j\Sigma_{k=1}^{s}\Sigma_{i=1}^{n}(G_{i,j,k})^2$, where $G_{i,j,k}$, defined in Eq. (2), was computed by substituting the volume measured in a phantom scan for $V_{2,i,j,k}$ and the corresponding true nodule volume for $V_{1,i,j,k}$. Any method that minimizes the sum of squared errors can be considered an optimal method.[27] For convenience, the rms error

$$E(j,n,s)=\sqrt{\frac{1}{ns}\sum_{k=1}^{s}\sum_{i=1}^{n}(G_{i,j,k})^2} \tag{3}$$

was computed for methods $j=1,\ldots,m$ to compare the performance of the m methods in estimating that no change in volume occurred.

2. Error measure on clinical data

The least-squares measure in Eq. (1) applied to the clinical data compared the growth estimate $G_{i,j}$ of a nodule i computed by method j with the growth estimate $G_i^{(r)}$ of this nodule provided by the radiologist. In particular,

$$\min_j\sum_{i=1}^{n}(G_{i,j}-G_i^{(r)})^2$$

$$=\min_j\sum_{i=1}^{n}\left(\frac{V_{2,i,j,k'}-V_{1,i,j,k}}{V_{1,i,j,k}}-\frac{V_{2,i,k'}^{(r)}-V_{1,i,k}^{(r)}}{V_{1,i,k}^{(r)}}\right)^2, \tag{4}$$

where the volume $V_{1,i,j,k}$ of a nodule i measured with method j in scan 1, taken with imaging parameters k, was compared to its volume $V_{2,i,j,k'}$ measured with method j in scan 2, taken with imaging parameters k'. Volumes $V_{2,i,k'}^{(r)}$ and $V_{1,i,k}^{(r)}$ were defined correspondingly.

We followed the analysis in Ref. 28 for comparison of observer studies to compare the radiologist's estimates with the automated methods. Assuming an approximately normal distribution of the difference in growth estimates, the 95% bounds of agreement between a method and the radiologist show the bounds between which 95% of the differences in measurements fall.

Radiologists have been shown to have bias and variance in their own measurements. The rms error might only reveal

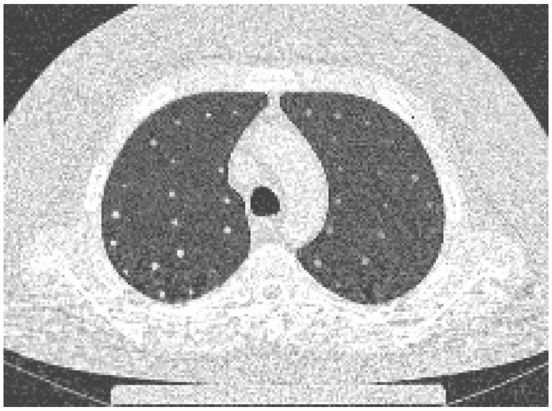

FIG. 3. CT scan of chest phantom.

TABLE II. Experimental data.

Patient	Scan	Kernel indicator of reconstruction algorithm	Pixel size (mm)	Exposure (mAs)	Number of nodules tested
1	First	B60f	0.66	120	2
1	Second	B60f	0.61	60	2
2	First	B70f	0.62	20	3
2	Second	B60f	0.62	120	4
3	First	B70f	0.57	20	2
3	Second	B60f	0.55	60	3
4	First	B50f	0.51	20	3
4	Second	B60f	0.52	100	2
5	First	B60f	0.68	120	5
5	Second	B60f	0.70	120	3
Phantom	B40f	B40f	0.74	20	40
Phantom	B60f	B60f	0.74	20	40

a bias that is readily accounted for in practice. The bounds of agreement are better measures of an automated method's performance than its rms error because they can both capture the bias and offer a means of comparison to the variance.

III. MATERIALS

A. Chest phantom

A 5-cm-thick chest phantom fabricated with plastic materials that simulate lung, muscle, fat, and bone when imaged on CT was used. This phantom was designed, constructed, measured, and imaged at New York University Medical Center.[10] The phantom contained 20 spherical plastic nodules in each lung (Fig. 3). Solid and ground-glass nodules were simulated by using two materials, composed of epoxy resins and urethanes, with respective specific gravities of 1.02 and 0.63 g/cm^3. Ko *et al.*[10] report that the attenuation of the phantom's lung parenchyma was -780 Hounsfield Units (HU), the attenuation of the plastic ground-glass nodules was approximately -360 HU, and of the solid nodules 50 HU. They obtained the nodule volumes by multiplying the specific gravity of the materials with the measured nodule weight. The volume measurements were determined to be accurate within 0.5%. Nodule volumes were 7.5, 18, 35, and 60.7 mm^3. Corresponding nodules diameters were 2.4, 3.2, 4.0, and 4.9 mm. Five solid and five ground-glass nodules of each size category were used. The nodules were separated from each other by at least 1 cm. Except for six of the nodules, three of each density, that were located adjacent to the lung wall, the nodules did not contact other structures.

In this work, the phantom was scanned four times. The phantom was moved after each scan was completed. Each scan was reconstructed both with a low-frequency (B40f) and high-frequency (B60f) reconstruction algorithm, resulting in 8 data sets. The phantom scans were taken on a multidetector row Siemens Somatom Volume Zoom Plus 4 CT using a 1 mm collimator for the entire study and were reconstructed in 1.25 mm sections at 1.0 mm increments using a 512×512 matrix (Table II). Images were quantized using 16 bits per pixel.

B. Clinical data

Five patients were selected from patients with thoracic CT scans taken for clinical evaluations at New York University. Each patient was evaluated in both an initial and a follow-up study. The studies occurred between January 2000 and September 2001. The number of days between scans was 268 on average. Studies were performed on a multi-detector row Siemens Somatom Volume Zoom Plus 4 CT in full inspiration using a 1 mm collimator and reconstructed with a 512×512 matrix in 1.25 mm sections at 1.0 mm intervals (Table II). A thoracic radiologist hand segmented 29 solid nodules. These hand segmentations were converted into nodule volumes (see Table III). Twenty of these were either isolated, which means that they were not connected to other structures in the lung such as blood vessels or lung fissures, or were adjacent to the pleural surface. None of these nodules were ground-glass nodules. Six of the nodules were less than 10 mm^3, seventeen were less than 60 mm^3. Six of the nodules larger than 60 mm^3 were attached to other structures. Figure 4 shows a CT image of a nodule used in testing.

IV. RESULTS

We tested the fixed-threshold, variable-threshold, and shape-based segmentation methods over a broad range of nodule variations on both phantom and clinical data.

A. Segmentation results for phantom data

On the phantom data, we tested $m=29$ variations of the segmentation methods on $n=40$ nodules in $s=8$ CT scan reconstructions. The fixed threshold method was tested for -700, -600, and -500 HU. These thresholds were also used as seed thresholds to compute the thresholds of the variable threshold methods. The shape-based methods were tested over the range -700 to -500 HU with 50 HU increments.

The performance of each method across all and with respect to particular nodule size and density variations was evaluated using Eq. (3). The performance of the methods

TABLE III. Volumes of nodules according to radiologist.

Patient	Nodule	Days between scans	Volume (mm^3) in initial scan	Volume (mm^3) in follow-up scan
1	1	294	118.76	122.24
1	2		28.33	22.57
2	1	294	8.25	12.70
2	2		19.89	31.74
2	3		28.63	4.88
2	4		a	58.59
3	1	394	57.75	42.69
3	2		a	8.85
3	3		3.21	3.85
4	1	116	381.98	203.39
4	2		5.80	a
4	3		40.62	31.57
5	1	241	328.28	a
5	2		640.79	a
5	3		61.92	161.35
5	4		100.47	286.45
5	5		67.17	166.19

[a] An accurate segmentation from the radiologist for this nodule instance was not obtained and so cannot be included in the growth estimation experiments.

TABLE IV. Root mean squared error by method on phantom data for all nodules where $E(j,n,s)=E(j,40,8)$ and for solid nodules where $E(j,n,s)=E(j,20,8)$.

Method	Rms error over all nodules	Rms error over solid nodules
2D shape	1.55	0.58
3D shape	0.54	0.32
$T_f=-700$ HU	9.79	9.52
$T_1^{(-700)}$	0.56	0.41
$T_2^{(-700)}$	1.83	1.91
$T_3^{(-700)}$	0.72	0.44
$T_4^{(-700)}$	0.63	0.75
$T_5^{(-700)}$	26.35	8.95
$T_6^{(-700)}$	31.81	16.36
$T_7^{(-700)}$	23.30	8.47
$T_8^{(-700)}$	23.91	9.39
$T_f=-600$ HU	1.05	1.38
$T_1^{(-600)}$	0.50	0.23
$T_2^{(-600)}$	0.60	0.80
$T_3^{(-600)}$	0.68	0.31
$T_4^{(-600)}$	0.43	0.29
$T_5^{(-600)}$	2.95	1.32
$T_6^{(-600)}$	3.15	1.56
$T_7^{(-600)}$	2.53	1.19
$T_8^{(-600)}$	2.54	1.22
$T_f=-500$ HU	0.49	0.55
$T_1^{(-500)}$	0.52	0.25
$T_2^{(-500)}$	0.49	0.55
$T_3^{(-500)}$	0.69	0.31
$T_4^{(-500)}$	0.51	0.25
$T_5^{(-500)}$	1.99	1.53
$T_6^{(-500)}$	2.04	1.59
$T_7^{(-500)}$	1.80	1.46
$T_8^{(-500)}$	1.80	1.46

overall and the performance with respect to solid nodules are shown as examples in Table IV.

The variable threshold method $T_4^{(-600)}$, computed with globally average lung and nodule values, is the overall best method with a root mean squared error of 0.43. The variable-threshold methods based on the global measures (except $T_1^{(-700)}$), the 3D shape-based method, and the fixed-threshold method with the highest threshold -500 HU performed at similar levels. Variable threshold method T_1, based on averaging the voxels with the globally lowest lung and highest nodule attenuation, performed consistently across all seed thresholds. The other global variable methods did not perform as well at the lowest seed threshold. Methods using a low fixed threshold and methods using local variable thresholds with low seed thresholds ($T^{(-700)}$ and $T^{(-600)}$) did not perform well in general.

The phantom tests showed that taking size and density properties into account when segmenting nodules yielded better results than the overall best method. For example, most segmentation methods, especially T_7, performed significantly better on larger nodules than on smaller nodules. The 3D shape-based segmentation performed the best on 7 and 18 mm^3 nodules with respective rms errors of 0.50 and 0.45.

For the variable threshold method, global thresholds were generally more accurate than local measures, however, local variable-threshold method $T_7^{(-600)}$ performed the best for 35 mm^3 nodules with a rms error of 0.18 and local variable threshold method $T_7^{(-700)}$ performed the best for 60.7 mm^3 nodules with a rms error of 0.14.

Most of the segmentation methods were more accurate on solid nodules than on groundglass nodules as should be expected because the attenuation values of solid nodules are more distinct from the attenuation values of lung tissue than the attenuation values of ground-glass nodules are. On solid nodules, variable threshold method $T_1^{(-600)}$, based on averaging the voxels with the globally lowest lung and highest

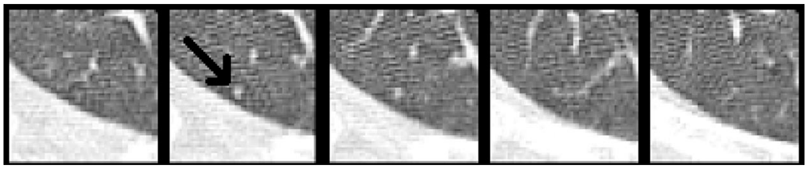

FIG. 4. Isolated nodule near lung border: This sequence is shown left to right from the top of the nodule.

TABLE V. Difference in volume from radiologist of selected methods on clinical data.

	Isolated nodules including nodules on lung surface			All nodules		
Method	Mean (mm^3)	Width/2 of 95% confidence interval (mm^3)	P value	Mean (mm^3)	Width/2 of 95% confidence interval (mm^3)	P value
3D shape	−17.3	12.7	0.010	13.8	47.1	0.55
T_f −600 HU	65.0	73.1	0.078	306.3	212.7	0.008
T_f −100 HU	−41.1	26.2	0.003	−37.3	29.3	0.017
$T_2^{(-600)}$	34.2	52.4	0.188	213.9	163.0	0.015
$T_2^{(-100)}$	−17.8	13.3	0.011	16.2	35.7	0.366
$T_5^{(-600)}$	6.8	28.1	0.619	99.9	92.7	0.040
$T_5^{(-100)}$	−14.9	18.4	0.107	5.9	29.6	0.690

nodule attenuation, performed the best with a root mean squared error of 0.23. On ground-glass nodules, variable-threshold method $T_2^{(-600)}$, based on the globally lowest lung and average nodule attenuation, performed the best with a root mean squared error of 0.25. Local measures performed especially poorly for ground-glass nodules. The smallest (7 and 18 mm^3) ground-glass nodules were the most difficult to segment. This is likely due to the partial volume effect having a much greater impact on these nodules than on larger, more dense nodules. In contrast, the larger ground-glass nodules were segmented with errors comparable to solid nodules.

Many of the methods performed better than the fixed threshold methods with statistical significance. For example, on solid nodules, the 3D shape-based method performed better than all the fixed threshold methods (all $p<0.001$). Global threshold method T_1 performed better than the fixed threshold methods at each corresponding seed threshold (all $p<0.001$). Local threshold method T_7 performed better than the corresponding fixed-threshold methods (all $p<0.001$) and better than the 3D shape-based method ($p<0.05$) on 60.7 mm^3 nodules. The difference between the performance of the methods on solid nodules in comparison to their performance on ground-glass nodules generally achieved statistical significance. The difference between the performance of many of the automated methods in comparison to other of the automated methods did not achieve statistical significance. Differences in performance in regards to reconstruction filters generally did not achieve statistical significance. For calculations of statistical significance we considered each of the eight reconstruction scans to be independent. If only four scans are considered independent, most of the differences in method performance retain the same degree of statistical significance. Of the above-mentioned examples, only the differences between method T_7 and the 3D shape-based method on 60.7 mm^3 nodules becomes statistically insignificant.

B. Segmentation results for clinical data

On clinical data, we tested $m=56$ variations of the segmentation methods on $n=12$ nodule pairs. We tested the fixed threshold method at 100 HU increments over the range −600 to −100 HU. We tested the variable threshold methods using all of these fixed thresholds as seed thresholds. We tested the shape-based methods over the same range with 50 HU increments. Note that the imaging parameters in initial and follow-up scans were generally not the same, i.e., $k \neq k'$. As we followed the analysis in Ref. 28, we first tested to see if the differences in volume and growth estimation between the methods and the radiologist were approximately normal. For growth analysis of all nodules and volumetric analysis of isolated nodules we found this to be true. For attached nodules the distribution of volume differences was not normal because of effects relating to nodule size.

The 3D shape-based method had the smallest 95% limits of agreement for isolated nodules with a mean difference in growth of -0.09 ± 0.33. Neither the 3D shape-based method nor any of the other methods produced statistically significant differences in growth estimates from the radiologist's estimates. Differences in many of the methods' *volume* estimates, however, were statistically significant. Table V presents volume differences for some of the methods tested.

V. DISCUSSION

Characterizing small nodules (<1 cm) as malignant has often relied on identifying nodule growth, most commonly expressed in terms of volume doubling time,[29] i.e., the time it takes a nodule to double its size. We can compare the errors in estimates of zero growth in phantom tests with this 100% point of reference. Consider two spherical nodules that are roughly the extremes of nodule sizes tested in the phantom. A one voxel increase in diameter for a nodule with a 6 voxel diameter indicates a 59% increase in volume (see Table I). A one voxel increase in diameter for a nodule with a 3 voxel diameter indicates a 137% increase in volume. Since the best method segmented nodules to within 43% of their actual size, this suggests that it was on average segmenting nodules to within one voxel of their surfaces and did not approach this 100% mark by grossly overestimating growth on average.

On the clinical data, the 95% bounds of agreement for growth for the majority of methods fell near the positive and negative 0.5 marks, which indicates the number of measurements that will be 50% larger or 50% smaller than the radiologist's segmentation. Consider a nodule that changed from 30 to 60 mm^3 as determined by a hand segmentation of the radiologist. An automated method that underestimated the growth by 50% might show this nodule's volume changing from 30 to 45 mm^3. Reliance on this automated method alone could significantly change the diagnosis of this nodule depending on the time between scans. Further testing needs to be done to determine the clinical significance of the automated methods' difference in agreement with the radiologist, especially since none of the methods produced statistically significant differences in growth estimates from the radiologist.

Using segmentations from additional radiologists or repeated segmentations by the same radiologist would provide a better model of what errors exist in current clinical practice by which to compare the automated methods. Such additional data would reveal whether human variations are larger than the difference between the average human observer estimates and the estimates by the automated methods.

Some of the error in the clinical tests can be attributed to the fact that the clinical images were not all taken with the same imaging parameters. A useful next step would be to conduct further testing in the clinical setting that included enough cases to isolate imaging parameters. The difference between growths calculated by the automated methods and those determined by the radiologist shows that more attention needs to be paid to accurate segmentation, especially of nodules connected to other structures in the lung. Notably, the consistently wide bounds of agreement with the radiologist for the fixed threshold methods and the generally poor performance of these methods on the phantom show that fixed threshold methods are inadequate. They would require too much specification to be useful under the wide variety of conditions that exist in a clinical setting.

The choice of the seed threshold did not effect all the methods in the same way. Some variable threshold methods were less sensitive to changes in the seed value than others, since these changes generally did not affect the maximum nodule and minimum lung values. It should also be noted that the described methods segment nodules by choosing a single threshold and therefore cannot separate nodules from adjacent vessels if the density of the contact point is higher than this threshold.

The global nodule measures drawn from the largest segmented object may be drawn from a large vessel structure and not from an actual nodule. The significance of this on the performance of the global threshold methods is unknown. To avoid the issue, a nodule detection algorithm could be used first and the segmentation methods could then be applied to the detected nodules only.

There was a difference between the seed thresholds selected for use in the phantom experiments and those used in the experiments on clinical data. Six seed thresholds starting with -600 HU were used for the clinical data, three thresholds starting with -700 HU were used for the phantom data. There were several reasons for this. First, because the phantom presented a known and fairly simple topology, experiments over a wider threshold range were not necessary. In the clinical data, the presence of other structures in the lung that could connect to nodules and imaging artifacts due to motion presented problems if a low threshold was selected as the seed threshold. From a few initial tests, $T_{s1} = -600\,\mathrm{HU}$ appeared to be a reasonable lower bound. This threshold vastly overestimates the size of nodules. This overestimation resulted in large 95% confidence intervals when results of the fixed threshold method $T_f = -600\,\mathrm{HU}$ were compared with the radiologist's segmentations as indicated in Table V. Conversely, we chose an upper bound $T_{sn} = 0\,\mathrm{HU}$ for the shape-based methods because large thresholds were likely to create very small objects that would satisfy the sphere occupancy constraint and produce an incorrect, drastically underestimated segmentation. This was especially true for nodules with strong connections to other structures. We only tested the shape-based method in increments of 50 HU. A smaller increment would provide somewhat better segmentations for some nodules, but would not greatly change the results of the experiments.

It should be noted that in some cases the shape-based segmentation methods produced multiple solutions from a nodule candidate when only a single nodule was present. We choose the best solution available of each nodule in computing the error. These multiple solutions were produced because an object segmented with a low threshold broke into several objects at a higher threshold. In practice this would require a physician to check the segmentations offered by the shape-based segmentation methods. The choices offered, however, would be significantly different from each other. Correct segmentations could be easily noted and incorrect segmentations rejected.

Figure 5 portrays in 2D the ideal case for segmentation. Here a nodule is spherical and the only change in segmentation of the nodule as the threshold increases is in the size of the segmented object and the strength of its contour. It is then reasonable to select the strongest contour as the best segmentation of that nodule. This ideal case represents a nodule with a fairly homogeneous density. Real nodules, however, can have a more heterogeneous density and are more difficult to segment. Figure 6 portrays a case in 2D in which a nodule is attached to another structure in the lung, a blood vessel perhaps. A low threshold, if it can separate the nodule from the larger vessel structure at all, segments a very spiculated object with a large amount of the blood vessel included. A higher threshold manages to remove most, but not all, of the blood vessel. At an even higher threshold, however, the nodule breaks into two pieces, one small and circular at the junction of the nodule and the vessel, the other a rough core of the nodule.

While Fig. 6 is a schematic illustration, it portrays a situation observed in the clinical data. There are many factors that may contribute to the occurrence of split objects, from imaging noise and motion artifacts to the presence of other lung structures. To avoid missing nodules or providing an incorrect segmentation when a better one is available, the method produced one segmentation result for each distinct

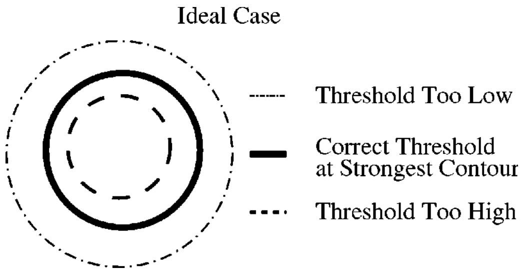

FIG. 5. The ideal nodule: As the segmenting threshold increases, the nodule shrinks around a central point. The strongest contour (shown with solid line) is chosen as the best segmentation.

piece of a candidate nodule that was found at a higher threshold.

In a few cases, the 2D shape-based method appears to be very accurate and vastly outperform its 3D variant. Such results were misleading because the voxels included in the nodule segmentation by the 2D method were very different from those segmented by the radiologist, especially for nodules connected to other structures in the lung. The erroneous segmentations produced by this method provided a compelling reason not to use 2D slice-based methods of examining nodules. Figure 7 presents a visual comparison of segmentations performed by the radiologist, by the 2D shape-based method, and by the 3D shape-based method for an isolated nodule, which resulted in somewhat similar, but clearly not identical segmentations. Figure 7 also presents a nodule connected to vessels in which the segmentations were all very different. The radiologist could more readily segment the central nodule connected to several vessels. The 3D shape-based method provided a poor segmentation that included both the nodules and the vessels. Even worse, the 2D shape-based method presented a useless disconnected collection of objects. This is because a single spiculated object may often have several disconnected protrusions in any single slice. Compactness measures are not well defined when treating separate objects as a single entity. Any method that could consistently piece together appropriate segmentations from collections of 2D objects is likely to be complex and any benefit gained from using simpler 2D methods will be lost. For shape-based methods, fully 3D measures are clearly needed.

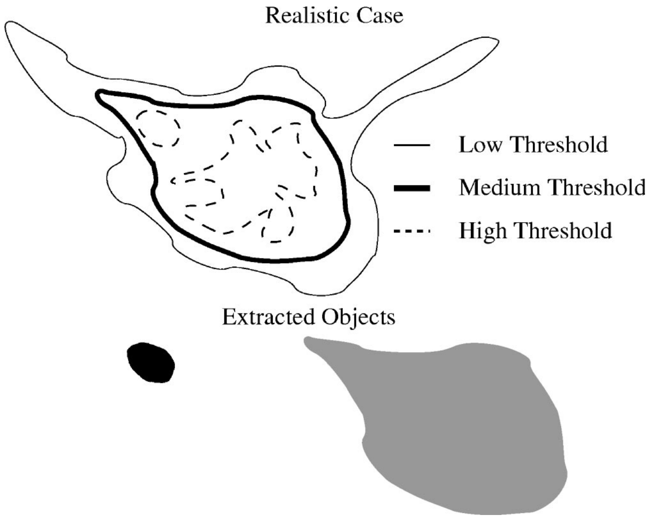

FIG. 6. The realistic nodule: Nodules often have complicated density structure. At a high threshold, this nodule breaks into two parts. The shape-based segmentation methods produce a segmentation for each of these parts.

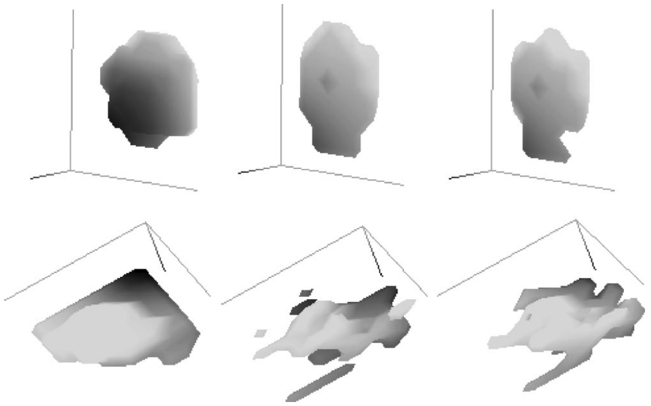

FIG. 7. Segmentation of isolated nodule (top) and nodule attached to vessels (bottom): radiologist (left), 2D method (middle), 3D method (right).

VI. CONCLUSIONS

We proposed several automated methods for segmenting nodules for the purposes of aiding physicians in the diagnosis of pulmonary metastasis of oncology patients and evaluation of the disease during treatment. Nodule segmentation and growth assessment are difficult tasks. Manual assessment is time consuming and prone to error. We showed that utilizing the full 3D nature of CT scans produced better segmentations than slice-by-slice examinations. No method achieved better than 0.43 rms error in volume measurement across all nodule variations on the phantom. However, when nodule size and density were examined separately, the 3D shape-based method and a few parametrizations of the variable threshold method showed strong improvement. Individual methods gave a rms error of 0.23 for solid nodules and 0.14 for nodules with volume 60.7 mm^3 on the phantom. On the clinical data the radiologist and the automated methods produced statistically significant differences in volume estimation, but not in growth estimation.

The ultimate goal is to design a growth estimation method that is at least as reliable as any radiologist. Toward this goal, future work would compare the growth estimates of several radiologists to the estimates of the automated methods. The contribution of the current work was to single out which methods perform best on specific nodule properties.

ACKNOWLEDGMENTS

Funding was provided by the National Science Foundation (IIS-0093367 and EIA-0202067), the Office of Naval Research (N000014-01-1-0444), and The Whitaker Foundation. Thanks to Sarah Dubauskas, Katelyn Mann, Harrison Hong, and Henry Rusinek of NYU for their support and Mark Crovella, Gene Itkis, Nick Makris, Eike Rietzel, and Shanghua Teng for helpful discussions.

[a]Electronic mail: mullally@cs.bu.edu

[1] Cancer Facts and Figures 2003, American Cancer Society. http://www.cancer.org, 2003.

[2] S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo, ''Cancer statistics, 1999,'' CA Cancer J. Clin. **49**, 8–31 (1999).

[3] C. I. Henschke, D. F. Yankelevitz, D. Libby, and M. Kimmel, ''CT screening for lung cancer: The first ten years,'' Cancer (N.Y.) **8**, 47–54 (2002).

[4] B. J. Hillman, ''Economic, legal, and ethical rationales for the ACRIN national lung screening trial of CT screening for lung cancer,'' Acad. Radiol. **10**, 349–350 (2003).

[5] D. P. Naidich, ''Helical computer tomography of the thorax,'' Radiol. Clin. North Am. **32**, 759–774 (1994).

[6] C. I. Henschke, D. I. McCauley, D. F. Yankelevitz, D. P. Naidich, G. McGuinness, O. S. Miettinen, D. M. Libby, M. W. Pasmantier, J. Koizumi, N. K. Altorki, and J. P. Smith, ''Early Lung Cancer Action Project: Overall design and findings from baseline screening,'' Lancet **354**, 99–105 (1999).

[7] D. F. Yankelevitz, R. Gupta, B. Zhao, and C. I. Henschke, ''Small pulmonary nodules: Evaluation with repeat CT—preliminary experience,'' Radiology **212**, 561–566 (1999).

[8] D. F. Yankelevitz, A. P. Reeves, W. J. Kostis, B. Zhao, and C. I. Henschke, ''Small pulmonary nodules: Volumetrically determined growth rates based on CT evaluation,'' Radiology **217**, 251–256 (2000).

[9] L. H. Schwartz, M. S. Ginsberg, D. DeCorato, L. N. Rothenberg, S. Einstein, P. Kijewski *et al.*, ''Evaluation of tumor measurements in oncology: Use of film-based and electronic techniques,'' J. Clin. Oncol. **18**, 2179–2184 (2000).

[10] J. P. Ko, H. Rusinek, E. Jacobs, R. Chandra, G. McGuinness, M. Betke, and D. P. Naidich, ''Volume quantitation of small pulmonary nodules on low dose chest CT: A phantom study,'' in Radiological Society of North America 87th Scientific Assembly and Annual Meeting (RSNA), Chicago, IL, November 2001.

[11] B. Zhao, D. Yankelevitz, A. Reeves, and C. I. Henschke, ''Two-dimensional multi-criterion segmentation of pulmonary nodules on helical CT images,'' Med. Phys. **26**, 889–895 (1999).

[12] B. Zhao, A. Reeves, D. Yankelevitz, and C. I. Henschke, ''Three-dimensional multicriterion automatic segmentation of pulmonary nodules of helical computed tomography images,'' Opt. Eng. (Bellingham) **38**, 1340–1347 (1999).

[13] B. Zhao, W. Kostis, A. Reeves, D. Yankelevitz, and C. Henschke, ''Consistent segmentation of repeat CT scans for growth assessment in pulmonary nodules,'' in Proceedings of the SPIE Conference on Image Processing, San Diego, CA, February 1999, Vol. 3661, pp. 1012–1018.

[14] P. F. Judy, F. L. Jacobson, B. Zhao, D. A. Israel, and C. Del Frate, ''CT lung nodule size: Effects of scanners and reconstruction filters,'' in Ref. 10.

[15] S. G. Armato, M. B. Altman, and P. J. La Rivière, ''Automated detection of lung nodules in CT scans: Effect of image reconstruction algorithm,'' Med. Phys. **30**, 461–472 (2003).

[16] S. G. Armato, M. L. Giger, C. J. Moran, J. T. Blackburn, K. Doi, and H. MacMahon, ''Computerized detection of pulmonary nodules on CT scans,'' Radiographics **19**, 1303–1311 (1999).

[17] M. S. Brown, M. F. McNitt-Gray, N. J. Mankovich, J. G. Goldin, J. Hiller, L. S. Wilson, and D. R. Aberle, ''Method for segmenting chest CT image data using an anatomical model: Preliminary results,'' IEEE Trans. Med. Imaging **16**, 828–839 (1997).

[18] L. Fan, J. Qian, G. Wei, C. Novak, and B. Odry, ''Automatic pulmonary nodule detection in multi-slice CT data,'' in Proceedings of the International Conference on Diagnostic Imaging and Analysis, Shanghai, China, August 2002, pp. 399–404.

[19] M. L. Giger, K. T. Bae, and H. MacMahon, ''Computerized detection of pulmonary nodules in computed tomography images,'' Invest. Radiol. **29**, 459–465 (1994).

[20] M. N. Gurcan, B. Sahiner, N. Petrick, H.-P. Chan, E. A. Kazerooni, P. N. Cascade, and L. Hadjiiski, ''Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system,'' Med. Phys. **29**, 2552–2558 (2002).

[21] J. P. Ko and M. Betke, ''Chest CT: Automated nodule detection and assessment of change over time—preliminary experience,'' Radiology **218**, 267–273 (2001).

[22] 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 Trans. Med. Imaging **20**, 595–604 (2001).

[23] W. Mullally, M. Betke, H. Hong, J. Wang, K. Mann, and J. P. Ko, ''Multicriterion 3D segmentation and registration of pulmonary nodules on CT: A preliminary investigation,'' in Ref. 18, pp. 176–181.

[24] Y. Kawata, N. Niki, H. Ohmatsu, M. Kusumoto, R. Kakinuma, K. Mori, H. Nishiyama, K. Eguchi, M. Kaneko, and N. Moriyama, ''Analysis of pulmonary nodule evolutions using a sequence of three-dimensional thoracic CT images,'' in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2001: Fourth International Conference, Utrecht, The Netherlands, October 2001*, edited by W. J. Niessen and M. A. Viergever (Springer, Berlin, 2001), pp. 103–110.

[25] R. Kostis, A. Reeves, D. Yankelevitz, and C. Henschke, ''Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images.'' IEEE Trans. Med. Imaging **22**, 1259–1274 (2003).

[26] B. K. P. Horn, *Robot Vision* (MIT, Cambridge, MA, 1986).

[27] S. M. Kay, *Statistical Signal Processing* (Prentice–Hall, Englewood Cliffs, NJ, 1993).

[28] D. G. Altman, *Practical Statistics for Medical Research* (Chapman and Hall/CRC, Boca Raton, FL, 1991).

[29] D. F. Yankelevitz and C. I. Henschke, ''Small solitary pulmonary nodules,'' Radiol. Clin. North Am. **38**, 471–478 (2000).