Synthetic Lung Tumor Data Sets for Comparison of Volumetric Algorithms[1]

Adele P. Peskin,[(a)(d)] Karen Kafadar, [(b)] Alden Dima,[(c)] Javier Bernal,[(c)] and David Gilsinn[(c)]

[(a)]National Institute of Standards and Technology, Boulder, Colorado 80305 USA
[(b)]Department of Statistics and Physics, Indiana University, Bloomington, Indiana 47408 USA
[(c)] National Institute of Standards and Technology, Gaithersburg, Maryland 20899 USA
[(d)]Corresponding author: peskin@boulder.nist.gov

Abstract **-** *We present a set of synthetic lung tumor data in which synthetic tumors of known volume are embedded in clinical lung computerized tomographic (CT) data in different background settings in the lung. Because the change in pulmonary nodules over time is an important indicator of lung tumor malignancy, it is important to be able to accurately measure changes in tumor size from measurements made at different times and possibly with different equipment. Standardized lung tumor data sets can be used to calibrate the differences between sets of scans as well as accurately compare volumetric measurement techniques. Our standard data sets combine the usefulness of phantom data with the clinical challenges of realistic CT scans.*

Keywords: image processing, segmentation, synthetic data, reference data.

1 Introduction

The change in pulmonary nodules over time is an extremely important indicator of tumor malignancy and rate of growth. Physicians base both diagnoses and treatment on perceived changes in size, so accurate and precise measurement of such changes can have significant implications for the patient. With current technology, tumor sizes, from which changes in size over time are calculated, are measured via computed tomography (CT), though often on different CT machines, with different operators, at different times of the day, and with patients in different physical positions relative to the CT equipment. Thus, a particular tumor is unlikely to be divided into slices at exactly the same places on two different sets of scans. The pixel distributions, intensity of grayscale, and average background values also may not be the same between two different sets of data.

Many articles have described a variety of techniques for calculating tumor volumes and/or the change in tumor size between two sets of data on the same tumor [1–6]. To compare these techniques, we need standardized data with known tumor volumes of various sizes within different levels of background noise. Although phantom tumor data are currently available and published studies [7,8] have compared volumetric methods on these phantom data sets, the phantom data settings are often not realistic, because the synthetic phantoms are placed in a synthetic background. A realistic and valid assessment of these volumetric methods needs realistic calibrated data sets, in which phantom ``tumors" of constructed size, shape, and volume are placed within data collected from clinical CT scans of background (non-tumor) tissue. These data sets should contain data that represent the different types of tumors seen in more realistic data. In this study we present a method for generating phantom lung tumors in various settings: centrally located in the lung and free of blood vessels, centrally located in the lung in a region where blood vessels also are located, and attached to the pleural surface of the lung by a tail. We present a set of synthetic lung tumor data in which synthetic tumors of known volume are embedded in clinical lung CT data for each of these situations. We show the resulting pixel distributions from many sets of clinical lung tumor data, and how we use these distributions to develop algorithms to create these data sets.

2 Objective

Our goal is to create sets of artificial tumors of known volumes embedded in real lung data, for use in comparing techniques that measure tumor size and growth. The data should include a wide variety of tumor sizes and shapes. Analyzing the artificial data should involve the same complications associated with making volume measurements of clinical tumors, such as those that arise due to connectivity to blood vessels and the pleural lining. The end result of our work is the creation of techniques to produce diverse types of

[1] This contribution of NIST, an agency of the U.S. Government, is not subject to copyright.

embedded tumors that cover the range of tumor types seen in real life, in terms of sizes, shapes, and confounding features, as needed to perform robust tests of volumetric software.

3 Data Description

To develop a method for creating standardized synthetic tumors, we studied the pixel distributions of many sets of images obtained from the Public Lung Database to Address Drug Response, which is funded by the Cancer Research and Prevention Foundation (www.via.cornell.edu/crpf.html). This database contains many examples of each of the types of tumors listed above. Figure 1a shows an example from this database of a single slice of data containing a tumor relatively free of blood vessels, and Figure 1b shows a region of the pixels in and around the tumor in this slice of data. To visualize the pixel distributions of the lung data sets, we discretize each individual pixel in each slice of data in the region of the tumor and color-code the categories accordingly. In this way, the clear differences in pixel intensities between pixels inside of lung tumors, pixels on or near the surfaces of lung tumors, and pixels in the surrounding lungs are apparent.

4 Data Creation

To create synthetic tumors embedded in CT data, we begin with a large, centrally located tumor, which is relatively free of blood vessels. The tumor is located in slices 110-150 of this data set, and for each slice we collect pixels in the region of the tumor. We are interested both in the pixel intensity distribution in this region as a whole, and in the sub-distributions at the tumor center and along the tumor edges. Figure 1c shows a histogram of all of the pixel intensities within the larger bounding box of Figure 1b, and are representative of the data sets in general. Figure 1d shows a histogram of all of the pixel intensities within the smaller bounding box shown in Figure 1b, which correspond to the interior of the tumor.

The actual edge of the tumors in CT data will not necessarily lie exactly along grid points, but it will lie within one pixel length of a grid point. Each pixel has a nominal location, but the intensity value at that pixel represents a volume in the original sample; we refer to this volume as a "voxel". The intensity of a pixel that is representing the edge of the tumor will vary with the distance between the closest pixel location and the actual edge. So the edge itself will not appear as a clear edge in the data, but rather as a blurred pixel-length region that represents the edge, and hence this region will have a distribution of pixel intensities much wider than that for the lung pixels. We attempt to define this distribution here.

We isolated the pixels in regions of many lung tumors from the Public Lung Database, and compared the pixel intensity distributions. We found that the mean and standard deviation of these distributions vary slightly across data sets and hence cannot be used as standard metrics. A more consistent metric across data sets comes from an empirical estimate of the magnitude of the gradient of the pixel intensities. The mean and standard deviation of the gradient magnitude at the tumor edges are consistent across a wide range of data sets [9]. This makes sense, because the lung tissue is a fairly consistent density, the tumor is a fairly consistent density, and the change between tumor and lung tissue covers a consistent range. The distribution of pixel intensities at the edge reflects this overall difference. The spread of intensities at the edge is in part a function of the physical spacing between adjacent pixels. If the locations are closer to one another, the edge will be more clearly defined. The pixel intensity associated with the highest point in the gradient of the intensity field, however, should not depend upon the pixel spacing.

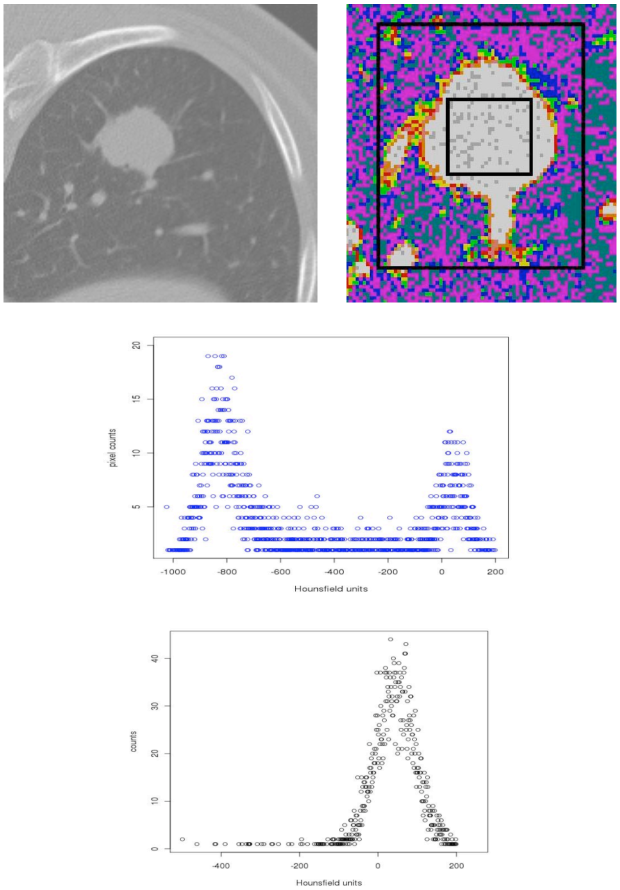

Figure 1: a.) Section of a slice of lung CT data containing a lung tumor; b.) Pixel intensities in Hounsfield units for data in 1a: white: -150 to 100, orange:-250 to -150, pink:-350 to -250, red :-450 t0 -350, yellow: -550 to -450, green: -650 to -550, blue: -750 to -650, purple: -850 to -750, teal: less than -850; c.) Histogram of the intensities inside the larger box of 1b; d.) Histogram of intensities inside smaller box of 1b.

We step back from the clinical data and use synthetic data to perform calibrations to determine exactly how to create our blurred boundary regions. Separately from the lung data, sets of synthetic images are created, containing spheres of sizes ranging from 4 to 20 pixel lengths. The pixel intensities at grid points in each of these grids represent the distance from the grid point to a central specified point in the grid. An isosurface within this grid at a particular value, therefore, represents an exact sphere with the radius of that isovalue. The goal is to represent a sphere of known volume with simulated data that has blurring at its boundary that is equivalent (or similar to) the blurring at edges of tumors in real CT scans. We compare the known analytical volumes of these spheres with volumes calculated from our technique to measure tumor volumes, described in [9], since we will use this measurement technique to analyze our new data. These results are shown in Table 1.

Table 1. Volumes of spheres measured with our marching cubes algorithm

radius	Calculated volume*	Analytical volume*	Percent error**
4	258.267	268.082	3.66
5	511.185	523.598	2.37
6	890.219	904.778	1.61
7	1420.494	1436.754	1.13
8	2124.761	2144.659	0.93
9	3031.675	3053.625	0.72
10	4164.123	4188.787	0.59
11	5547.871	5575.275	0.49
12	7208.727	7238.223	0.41
13	9170.843	9202.765	0.35
14	11460.105	11494.030	0.30
15	14099.811	14137.155	0.26
16	17118.117	17157.270	0.23
17	20537.883	20579.510	0.20
18	24385.213	24429.004	0.18
19	28684.250	28730.889	0.16
20	33460.980	33510.293	0.15

(*cubic pixel length)(**100(analytical-caculated)/analytical))

Now we create similar spheres of each size, but the edges of these spheres will represent the blurred edges of the clinical images. Instead of using distances from a central point for the pixel intensities in our synthetic data sets, we now use clinical intensities sampled from lung tumor data. All pixel intensities are taken from the tumor of the data set shown in Figure 1. Pixel intensities for pixel locations that are entirely inside of the sphere are chosen at random from a collection of intensities representing the clinical tumor data, shown in Figure 1d. Pixel intensities for all pixels on or near the edge are calculated differently.

Each pixel location is defined as the center point of a unit sized voxel. The intensity of that pixel depends on the density not just at the pixel location, but the density throughout that voxel. We compute a weighted average of subsamples where the weight is dependent on the distance to the center of the voxel and the subsample points are evenly distributed through the voxel. We divide the voxel into 100 x 100 x 100 pieces. If a piece is outside of the sphere, it contributes nothing. If it is inside the sphere, it contributes a term that is proportional to the inverse of its distance from the central pixel location. In the case of a voxel entirely within the sphere, this sum is equal to 1.0. For a voxel at the edge, this sum is a fraction, *f*, between 0.0 and 1.0. The actual value of the pixel intensity for the center point of the voxel is then:

$$\begin{aligned} &\mathit{Intensity} = f * (\mathit{tumor\ mean\ value}) + \\ &(1.0\text{-}f) * (\mathit{maximum\ background\ value}) \end{aligned} \tag{1}$$

The maximum background value of pixel intensity in the region of the tumor is defined as a constant, k2, for a particular data set, as described in [9]. The images of spheres with blurred edges are then embedded into the slices of the data set. Figure 2a is a section of a slice of the resulting data. We then calculate the magnitude of the new gradient of the pixel intensity field and compare it to the gradient of the pixel intensity field in the region of the clinical tumor in that data set. For each tumor, the clinical tumor and the synthetic sphere, we find the average magnitude of the gradient at each intensity value between -800 Hounsfield units (HU) and -100 HU. These curves are smoothed using a locally weighted linear smoother (using R software; see www.r-project.org), and presented in Figure 2b, along with a graph of the normalized differences in the square roots of the gradient magnitudes from the two tumors, Figure 2c. These figures

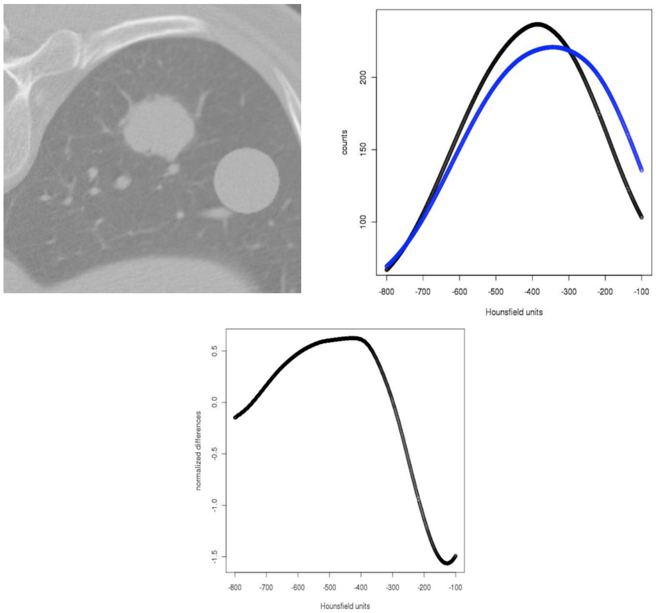

Figure 2: a.) Section of a slice of lung data with embedded sphere; b.) Smoothed curves of the average gradient magnitude of the pixel intensity in the region of the clinical tumor (in black) and the synthetic sphere (in blue); c.) Normalized differences in the square roots of the curves in 2b.

show the similarity between the gradients of the pixel intensity fields of the clinical tumor and the synthetic tumors.

5 TYPES OF LUNG DATA

Sets of synthetic spheres with blurred edges were then embedded into a section of the lung data where clusters of blood vessels reside. Where the blood vessels and sphere edges overlap, the edge of the sphere blends with the pixels representing the blood vessels. Figure 3a shows a spherical synthetic tumor embedded into a section of the lung surrounded by blood vessels. This is an example of a data set that could be used to evaluate methods that eliminate attached blood vessels from tumor volume measurements. By situating geometric objects of known volume into regions of lung tissue in areas that lead to difficult volumetric measurements, we can study and compare methods that deal with removing vascular attachments from tumor volume measurements. The edges of these objects will be realistic blurred edges, constructed to maintain the volumes associated with objects with precise edges.

Attaching synthetic tumors to the pleural lining of lung tumor data can be done in many different ways. Most of the tumor data we have investigated so far has shown that the extent of attachment of the tumors varies over the slices containing the tumor. In each we have found, by looking at 3D pictures of these tumors and their attachments, slices in which the tumor pixels blend into the pixels of the lining, and neighboring slices in which the attachment gradually disappears. In many of these examples, networks of blood vessels also complicate the distribution of middle range edge pixel values.

An example of a synthetic sphere that is attached to the pleural lining is given in Figure 3b. A geometric attachment is created, whose pixels are selected at random from the tumor in this data set, and the pixel intensities at the edges of the attachment are calculated according to our integration method. The extent of attachment of the sphere to the pleural lining varies in different slices containing the sphere, and tapers off as in seen in each example of the clinical data.

We investigated one last example of this type of synthetic data, shown in Figure 3c. Micro CT data of a phantom tumor from the FDA, pictured in Figure 3d, is embedded into the lung data. As in the synthetic spherical tumors, the edges of the phantom tumor cover a one pixel length edge, in which the pixel intensities vary with the distance from the edge of the tumor. Table 2 gives the embedded geometries and placements of the various synthetic data sets we have created so far.

5 Volumetric Comparisons

Clinical lung CT data of lung tumors do not have well-defined edges, and the strength of a method for determining lung tumor volumes from these data depends on how well it defines those edges. We use the synthetic lung tumor data to begin to evaluate a variety of methods for lung tumor volume measurement. Estimates of the embedded sphere volumes for the case of a radius of 20 pixel lengths show clearly that the volume is highly dependent on the method of determining the sphere boundary. The pixel spacing within a slice is 0.57 mm, and 1.25 mm between slices, i.e., the sphere has a radius of 11.4 mm and a volume of 6205.87 cubic mm. The volume estimate by the method described in [9] was 6196.74 cubic mm or approximately a 0.15 % error. The radius estimate in this case would be 11.39 mm. Thus a 0.09 % radius error produced a 0.15 % volume error. A second method based on a Canny edge detection algorithm produced a volume of 6596.88 cubic mm with a radius of 11.63 mm. In this case a 2 % error in radius estimate produced a 6.3 % volume error. A third method involving approximately the edge produced by a Canny edge algorithm with B-splines gave a volume of 6626.57 cubic mm was based on a radius estimate of 11.65 mm. In this case a 2.2 % error estimate for the radius produced a 6.8 % volumetric error. We also computed a volume using a commercial software package using several accurate but very interactive methods of region selection and refinement. Depending upon the specific set of steps selected

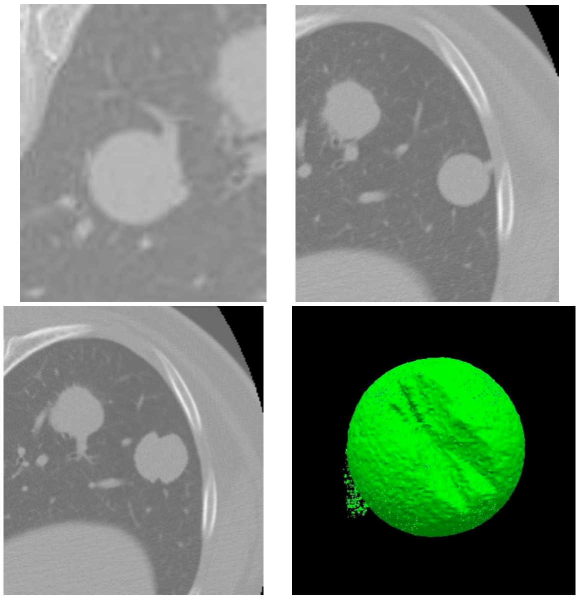

Figure 3: a.) Section of a slice of lung data with a sphere embedded in the location of a large blood vessel; b.) Section of a slice of lung data with a sphere attached to the pleural lining of the lung; c.) Section of a slice of lung data with a phantom tumor embedded; d.) Isosurface at 30 HU of the phantom tumor.

for the measurement, the volumetric accuracy varied between 1.13 % and 2.95 % error. Clearly the method to approximate a tumor boundary requires a great deal of accuracy to lead to good volume measurements. These data sets now provide a tool to compare different types of measurement strategies.

Table 2. Sets of available synthetic tumor data

geometry	Sphere radius*	Ellipsoid dimensions	type**
sphere	10		free
sphere	15		free
sphere	20		free
sphere	15		embedded
sphere	15		attached
ellipsoid		20-10-10	free
ellipsoid		10-20-10	free
ellipsoid		10-10-20	free
ellipsoid		10-20-10 rotated 45 degrees about slice direction	free
ellipsoid		10-10-20 rotated 45 degrees perpendicular to slice	free
FDA	phantom		

*in pixel lengths, system variant

**free = free of blood vessels, embedded = embedded in vessels, attached = attached to pleural lining

6 Conclusions and Future Work

We have created an initial group of synthetic lung tumor data sets, in which both known geometric shapes and well defined phantom tumors have been embedded into clinical lung tumor data. We have developed methods for the insertion of these synthetic tumors in different regions of the lung data, so that the synthetic tumors can be used as standardized data to test methods for measuring lung tumors whose edges are either clearly defined or partially hidden by local blood vessels in the lung. The methods proposed here will allow researchers to create their own synthetic sets, enabling a systematic comparison of the volumetric methods currently available to measure tumor size and growth. Lung tumors often grow in non-symmetric directions, in the shapes of spines and knobs growing from an original sphere-like region. Future work in the creation of these synthetic data sets includes plans to create a variety of characteristic synthetic shapes in the lung tumor data, and to greatly expand the sets of data we currently have available.

ACKNOWLEDGEMENTS

Support for Dr. Kafadar from the Army Research Office, Grant Number W911NF-05-1-0490 awarded to the University of Colorado-Denver, and from the National Science Foundation, Grant Number DMS0802295 awarded to Indiana University, is gratefully acknowledged.

7 References

[1]Kostis, W.J., Reeves, A.P., Yankelevitz, D.F., Henschke, C.I. Three-Dimensional Segmentation and Growth-Rate Estimation of Small Pulmonary Nodules in Helical CT Images. *IEEE Trans. on Medical Imaging* **22**, No. 10, October 2003.

[2]Reeves, A.P., Chan, A. B., Yankelevitz, D.F., Henschke, Kressler, C.I.B., Kostis, W.J. On measuring the change in size of pulmonary nodules. *IEEE Trans. on Medical Imaging* IEEE **25**(4):435-450 (2006).

[3]Mendonca., P., Bhotika, R., Sirohey, S., Turner, W., Miller, J., Avila, R.S. Model-based Analysis of Local Shape for Lesion Detection in CT Lung Images. *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2005)*. October 2005.

[4]McCulloch,C.C., Kaucic, R.A., Mendonca, P.R., Walter, D.J., Avila, R.S. Model-based Detection of Lung Nodules in Computed Tomography Exams. *Academic Radiology*. March 2004.

[5]Preim, B., Bartz, D. Image Analysis for Medical Visualization. *Visualization in Medicine* 83-131, 2007.

[6]Preim,B., Bartz, D. Exploration of Dynamic Medical Volume Data. *Visualization in Medicine* 83-131, 2007.

[7]Das, M., Ley-Zaporozhan, J., Gietema, H.A., Czech, A., Nuhlenbruch, G., Mahnken, A.H., Katoh, M., Bakai, A., Salganicoff, M., Diederich, S., Prokop, M., Kauczor, H., Gunther, R.W., Wildberger, J.E. Accuracy of automated volumetry of pulmonary nodules across different multislice CT scanners. *Eur Radiol* **17:**1979-1984 (2007).

[8]Ko, J.P., Rusinek, H.,Jacobs, E.L., Babb, J.S., Betke, M., McGuinness, G., Naidich, D.P.: Small Pulmonary Nodules: Volume Measurement at Chest CT-Phantom Study. *Radiology* **228**:864-70 (2003).

[9]Peskin, A.P., Kafadar, K.,Santos, A.M., Haemer, G.G. Robust Volume Calculations of Tumors of Various Sizes. *2009 International Conference on Image Processing, Computer Vision, and Pattern Recognition*.