Evaluation of reproducibility for manual and semi-automated feature extraction in CT and MR images

Edward A. Ashton, Larry Molinelli*, Saara Totterman, Kevin J. Parker**
Dept. of Radiology, **Dept. of Electrical Engineering, University of Rochester, Rochester, NY
*VirtualScopics, LLC, Rochester, NY
edashton@ece.rochester.edu

Abstract

Three methods for extraction and quantitative measurement of features in CT and MR images are examined: hand tracing, semi-automated tracing using the livewire graph search algorithm, and extraction using a geometrically constrained region growth algorithm. Extracted structures are evaluated in terms of volume, cross-sectional area, and major axis in plane. Reproducibility, time required for extraction, and accuracy of each of these metrics is measured for each of the extraction methods using both phantoms and clinical lung tumor data.

1. Introduction

The accurate measurement of soft tissue lesions and the tracking of small changes in lesions over time are topics of great interest both to pharmaceutical companies conducting drug trials and to clinicians attempting to monitor disease progression. Standard practice for obtaining such measurements currently involves either direct measurement taken from films or computer-aided manual tracing of lesion borders. Both of these processes are time consuming and prone to both inter- and intra-operator variation. In this paper we evaluate two potential techniques for automating this process, in the hope that these methods will provide improvements over manual measurement in terms of reproducibility, accuracy, and time required for measurement.

The first method evaluated is the well-known LiveWire[1,2] algorithm. This technique models the image under consideration as a weighted graph, with the weights to travel from each pixel to its neighbors defined by some desired image property. In our implementation, weights are set using both edge information and grayscale similarity. An optimal path is then calculated between user-defined points using a modified Dijkstra's shortest path algorithm.

The second automated method, geometrically constrained region growth (GEORG) combines elements of competitive region growth [3,4] and deformable

template [5,6] techniques, and requires human interaction only for the initial location of the lesion centroid. GEORG is a generalized technique that can make use of many different kinds of image information. In other work, GEORG has been studied using grayscale segmentation [7], competitive region growth [3] and multivariate maximum likelihood [8] as metrics for determining region boundaries. In this paper, we examine the performance of this algorithm using edge information in combination with a flexible, user-defined geometric model.

2. Measurement Techniques

The first semi-automated technique under consideration, LiveWire, uses a weighted graph search algorithm to connect points spaced around the perimeter of the object to be extracted. The implementation in this work made use of an edge-list graph implementation and a modified Dijkstra's shortest path algorithm. Weights between pixels at points i and j were given by:

$$w_{i,j} = abs(m - v_j) - e_j \qquad (1)$$

where m is the mean value of the end points, v_j is the grayscale value at pixel j, and e_j is the edge-map value at pixel j. Values in the original grayscale image and in the edge-map were normalized to prevent negative weights. Edge information was calculated in this implementation by application of a median filter of support 5x5, followed by a Laplacian filter of support 3x3. Operation of this algorithm is illustrated in Figure 1.

The second semi-automated method examined was GEORG. This technique requires a user to place a seed or string of seeds within each desired structure throughout the volume using one or more mouse clicks. The seed regions then expand into neighboring voxels provided that two constraints are satisfied: the edge information contained in the neighboring voxel must have a high probability of falling within the statistical distribution defined by all current included voxels, and inclusion of the neighboring voxel must not cause the

shape of the included region to deviate excessively from the *a priori* regional shape model. It is the first constraint that distinguishes this approach from deformable template techniques such as that described by Carlbom *et al.*[5], and the second constraint that distinguishes it from competitive region growth algorithms such as that described by Taylor and Barrett [4]. In this implementation *a priori* model is defined by the distribution of seeds within the structure. Once initiated, the expansion process continues until a stable boundary has been established. The operation of this algorithm is illustrated in Figure 2.
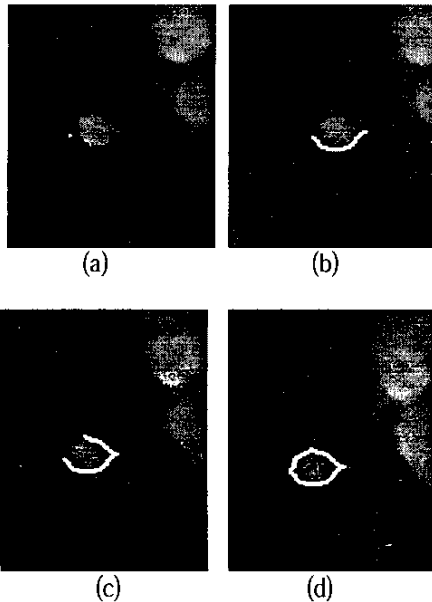

(a)   (b)

(c)   (d)

Figure 1: The lesion shown in this series is identified using the LiveWire algorithm in three mouse clicks. Connections between end-points are made using Dijkstra's shortest path algorithm.

3. Experimental Procedure

The experiments involved in this study were intended to assess the performance of the two algorithms under consideration with respect to manual in terms of three parameters: speed, precision, and accuracy. Two data sets, one clinical and one phantom-based, were used to assess required processing time for each algorithm, inter- and intra-operator measurement variability, and global accuracy.
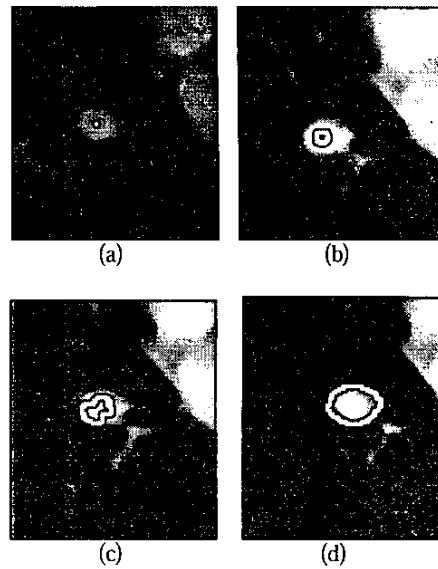

(a)   (b)

(c)   (d)

Figure 2: The lesion shown in this series is identified using the GEORG algorithm with a single mouse click. A seed is placed by the user in (a). Contour growth continues until a stable border is identified in (d).

3.1 Experimental data

The first experiment in this study was intended to assess the accuracy of all three measurement techniques in a reliable and objective manner. To that end, a magnetic resonance phantom intended to mimic abdominal organ tumors was obtained. The phantom specifications included the precise size and volume for each simulated tumor. This phantom was scanned in seven different orientations using a General Electric 1.5 Tesla Horizon MR Scanner. Sample images from this data set are shown in Figure 3.
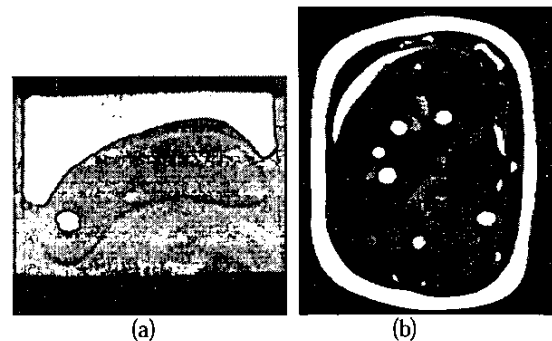

(a)   (b)

Figure 3: Sample images from two of the seven MRI organ tumor phantom data sets.

The second experiment in this study was intended to assess the precision and ease of use of each of the three methods under examination, as well as to examine their applicability to real-world data. To that end, clinical CT data for 15 small-cell lung cancer patients were obtained from the University of Rochester Medical Center. Each patient had between one and eight repeat studies over a period of 18 months, for a total of 75 CT volumes. Sample images from this data set are shown in Figure 4.
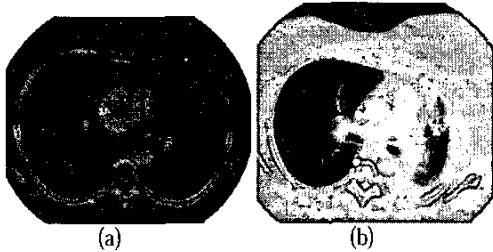

(a)   (b)

Figure 4: Sample images from two of the 75 CT lung cancer studies.

3.2 Experiments using phantom data

The phantom data sets allowed experiments designed to determine absolute accuracy for both manual and semi-automated techniques. Three distinct experiments were carried out using this data. First, each phantom data set was evaluated once using GEORG, once using manual tracing, and once using LiveWire. These results were compared to manufacturer's specifications for lesion size and placement. Second, each phantom data set was evaluated in seven separate trials by a single operator using each of the three measurement techniques. This allowed an estimate of the intra-operator variability associated with each method. In addition, comparison of these results to those of the previous experiment allowed an estimation of the variance introduced by the changes in orientation in the scanning process.

3.3 Experiments using clinical data

The experimental procedure used with the clinical data was very similar to that of the phantom data. All 75 clinical data sets were evaluated once using manual tracing, once using GEORG, and once using LiveWire. A sub-set consisting of 15 studies taken from five patients was then evaluated multiple times by single operators and once by four separate operators using each measurement method, allowing an estimation of intra- and inter-operator variability for each technique in a clinical setting.

4. Experimental Results

4.1 Results for phantom experiment

Results of the intra-operator variability study involving multiple phantom orientations are given in Table 1.

	Trace	Live Wire	GEORG
O1	69.9	59.6	59.0
O2	72.5	62.8	61.5
O3	66.2	60.2	64.0
O4	66.7	62.2	62.7
O5	68.9	61.7	63.8
O6	74.4	61.4	61.9
O7	68.1	60.4	60.5
av	69.5	61.2	61.9
sd	3.0	1.2	1.8
cv	4.3%	2.0%	2.9%
mt	30m	10m	10m

Table 1: Measurement of phantom lesion burden in cubic centimeters for seven scans, each with a different orientation. True lesion burden in this case was 65.2 cc. In the final four rows, av = average volume, sd = standard deviation, cv = coefficient of variation, and mt = mean analyisis time.

These results illustrate two basic points. First, both GEORG and LiveWire provide significant improvement over manual tracing in terms of operator time required and measurement precision. Second, both LiveWire and GEORG provide results which are somewhat closer to the objective truth than manual tracing, with GEORG and LiveWire under-estimating true lesion burden by 5.1% and 6.1%, respectively, while manual tracing over-estimates true lesion burden by 6.6%. This result is predictable. Manual tracers tend to draw outside the boundary of a given lesion in an effort to ensure that the entire lesion is identified, while both GEORG and LiveWire tend to find boundaries which are at the periphery of, but included in the lesion, and therefore to miss some pixels around the lesion rim.

4.2 Results for clinical experiments

The first question that must be answered for any study involving human clinical data relates to absolute accuracy. In short, is there any way to achieve an acceptable gold standard result? In many studies experimental results are compared to manual tracings as a proxy for absolute accuracy. To test the viability of that approach for clinical lung cancer data, we provided

four experienced radiologists with data for four patients from our clinical data set and asked them to trace all lung lesions. Total lesion burden results for this experiment are given in Table 2.

	Pat. 1	Pat. 2	Pat. 3	Pat. 4
Obs. 1	7394	15068	6764	25169
Obs. 2	12688	21435	10267	47114
Obs. 3	15907	33121	10649	49883
Obs. 4	18679	24948	15194	48615
Mean	13667	23643	10718	42695
CV	35.4%	31.8%	32.3%	27.5%

Table 2: Results of manual inter-operator variability experiment. Lesion burdens are given in cubic millimeters. Clearly, this level of precision does not support confidence in tracing as a gold standard.

Because manual tracing could not provide an accurate estimate of true lesion burden in these experiments, the only parameter that could be measured with confidence for the automated processes was precision. To this end, we asked four trained observers to measure total lesion burden for seven patients using both LiveWire and GEORG. Results for this experiment are given in Tables 3 and 4.

	Obs.1	Obs.2	Obs.3	Obs.4	Mean	CV
Pat. 1	23.7	24.9	24.6	23.8	24.3	2.4%
Pat. 2	3.53	3.55	3.53	3.68	3.57	2.1%
Pat. 3	29.3	28.3	30.0	28.1	28.9	3.1%
Pat. 4	5.66	6.79	6.29	6.06	6.20	7.6%
Pat. 5	36.2	33.2	37.8	38.8	36.5	6.7%
Pat. 6	12.3	12.4	12.8	12.5	12.5	1.7%
Pat. 7	7.20	7.61	7.21	6.70	7.18	5.2%

Table 3: Results for clinical inter-operator variability study using LiveWire. Volumes are given in cubic centimeters. Note that precision is roughly an order of magnitude better than that provided by manual tracing.

5. Conclusions

In general, these experiments indicate that both semi-automated methods provide a significant advantage over manual tracing in terms of both speed and precision. Additionally, the phantom experiments indicate that these methods are comparable to tracing in terms of absolute accuracy. Future work will include development of more realistic accuracy assessments.

	Obs.1	Obs.2	Obs.3	Obs.4	Mean	CV
Pat. 1	22.7	24.5	21.5	23.3	23.0	5.4%
Pat. 2	3.45	3.53	3.14	3.07	3.30	6.9%
Pat. 3	26.4	26.3	27.0	24.4	26.0	4.4%
Pat. 4	6.12	5.57	5.71	5.84	5.81	4.0%
Pat. 5	34.1	37.3	33.6	33.9	34.7	4.9%
Pat. 6	12.1	11.7	11.3	11.8	11.7	2.9%
Pat. 7	6.25	6.70	6.56	6.82	6.58	3.7%

Table 4: Results for clinical inter-operator variability study using GEORG. Precision is comparable to that for LiveWire. Mean results are consistently marginally lower than those provided by LiveWire.

References

[1] W. Barrett and E. Mortensen, "Interactive live-wire boundary extraction," Medical Image Analysis 1(4), pp. 331 – 341, 1997.

[2] A. Falcao, J. Udupa, and F. Miyazawa, "An ultra-fast user-steered image segmentation paradigm: Live wire on the fly," IEEE Trans. Medical Imaging 19(1), pp. 55 – 61, 2000.

[3] E. Ashton, S. Totterman, C. Takahashi, J. Tamez-Pena, and K. Parker, "Automated measurement of structures in CT and MR imagery: A validation study," Proc. 14th IEEE Symposium on Computer-Based Medical Systems, pp. 300 – 306, 2001.

[4] D. Taylor and W. Barrett, "Image segmentation using globally optimum growth in three dimensions with an adaptive feature set," Visualization in Biomedical Computing, pp. 98 – 107, 1994.

[5] I. Carlbom, D. Terzopoulos, and K. Harris, "Computer assisted registration, segmentation and 3D reconstruction from images of neuronal tissue sections," IEEE Trans. Medical Imaging 13, pp. 351 – 362, 1994.

[6] R. Chung and C. Ho, "3-D reconstruction from tomographic data using 2-D active contours" Computers and Biomedical Research, vol. 33, pp. 186 – 210, 2000.

[7] E. Ashton, K. Parker, M. Berg, and C. Chen, "A novel volumetric feature extraction technique, with applications to MR images," IEEE Trans. Medical Imaging 16(4), pp. 365 – 371, 1997.

[8] E. Ashton, C. Takahashi, M. Berg, A. Goodman, S. Totterman, "Automated quantification of MS lesions in MRI: A validation study," Proc. of SPIE-Medical Imaging, 2002.