Comparison of two-dimensional and three-dimensional iterative watershed segmentation methods in hepatic tumor volumetrics

Shonket Ray^{a)}

Department of Biomedical Engineering, University of California, Davis, California, One Shields Avenue, Davis, California 95616 and Department of Radiology, U.C. Davis Medical Center, 4860 Y Street, Sacramento, California 95817

Rosalie Hagge, Marijo Gillen, Miguel Cerejo, and Shidrokh Shakeri

Department of Radiology, U.C. Davis Medical Center, 4860 Y Street, Sacramento, California 95817

Laurel Beckett and Tamara Greasby

Department of Public Health Sciences, Division of Biostatistics, University of California, Davis, One Shields Avenue, Davis, California 95616

Ramsey D. Badawi

Department of Biomedical Engineering, University of California, Davis, California, One Shields Avenue, Davis, California 95616 and Department of Radiology, U.C. Davis Medical Center, 4860 Y Street, Sacramento, California 95817

(Received 7 April 2008; accepted for publication 7 October 2008; published 24 November 2008)

In this work the authors compare the accuracy of two-dimensional (2D) and three-dimensional (3D) implementations of a computer-aided image segmentation method to that of physician observers (using manual outlining) for volume measurements of liver tumors visualized with diagnostic contrast-enhanced and PET/CT-based non-contrast-enhanced (PET-CT) CT scans. The method assessed is a hybridization of the watershed method using observer-set markers with a gradient vector flow approach. This method is known as the iterative watershed segmentation (IWS) method. Initial assessments are performed using software phantoms that model a range of tumor shapes, noise levels, and noise qualities. IWS is then applied to CT image sets of patients with identified hepatic tumors and compared to the physicians' manual outlines on the same tumors. The repeatability of the physicians' measurements is also assessed. IWS utilizes multiple levels of segmentation performed with the use of "fuzzy regions" that could be considered part of a selected tumor. In phantom studies, the outermost volume outline for level 1 (called level 1_1 consisting of inner region plus fuzzy region) was generally the most accurate. For *in vivo* studies, the level 1_1 and the second outermost outline for level 2 (called level 2_2 consisting of inner region plus two fuzzy regions) typically had the smallest percent error values when compared to physician observer volume estimates. Our data indicate that allowing the operator to choose the "best result" level iteration outline from all generated outlines would likely give the more accurate volume for a given tumor rather than automatically choosing a particular level iteration outline. The preliminary *in vivo* results indicate that 2D-IWS is likely to be more accurate than 3D-IWS in relation to the observer volume estimates. © 2008 American Association of Physicists in Medicine.

[DOI: [10.1118/1.3013561](https://doi.org/10.1118/1.3013561)]

Key words: Tumor segmentation, iterative watershed method, marker-based watershed, RECIST, gradient vector flow, active contours

I. INTRODUCTION

Changes in tumor size as assessed by radiological imaging are widely recognized as a surrogate marker for response to anticancer drug therapies.^{1,2} In recent years, the response evaluation criteria in solid tumors (RECIST) (Ref. 3) have been adopted by a range of organizations and institutions in the field and have become the *de facto* standard. Briefly summarized, the RECIST methodology is a four-state classification of target lesions in terms of size change (determined using an anatomical imaging modality such as CT or MR) in response to a time table of therapy. Tumor size is measured using unidimensional measurements, which consist of calculating the sum of the longest diameters (SLDs) of target le-

sions with the initial baseline scan SLD measurement used as a reference point.

One of the main limitations of unidimensional methodology is poor retest repeatability. This is due both to difficulties in manual determination of the longest diameter and to possible orientation shifts, where the unidimensional measurement change in a target lesion might be incorrect due to differing image scanning orientations with respect to the tumor. Since tumor shapes are often nonspherical, the SLD measurement could change greatly and this can affect the overall tumor size change measurement. In addition, the response might be nonisotropic, reducing the sensitivity of the approach.

It has been hypothesized that such problems might be ameliorated by utilizing bidimensional measurements—these consist of sums of the products of the two longest perpendicular diameters.⁴ Some studies have further extended this concept to tridimensional measurements. However, when used to segment populations into RECIST-type response groups, there is little evidence that these approaches improve clinical utility.⁵ It is not clear whether this is due to image processing limitations or due to limitations of the RECIST cut points. These are based on nonimaging size-estimation accuracy criteria rather than the biology of cancer and cancer therapy.⁶

Recently, a range of anticancer agents have been developed with cytostatic rather than cytotoxic properties.⁷ These include receptor tyrosine kinase inhibitors such as imatinib (GleevecTM)⁸ and gefitinib (IressaTM).⁹ For such agents, tumor size change is not as rapid as it is in conventional cytotoxic drug therapy. This reduction in rate of size change both exacerbates the problems associated with unidimensional measurement error and further calls into question the four-state classifier associated with the RECIST methodology since imprecise measures of tumor dimensions and inappropriate classifications may lead to incorrect assessment of treatment efficacy.¹⁰

Test repeatability may potentially be improved by determining tumor volume rather than SLD.¹¹ This has been previously used with some success in studies involving neurological disorders¹² and in the lung^{13,14} leading to the notion that this specific metric can be extended for use in other parts of the body (for example, in oral cavity carcinoma¹⁵).

Tumor volumes may be determined in a straightforward manner by manual segmentation of tumor from normal tissue in a series of contiguous image slices but this process needs to be performed by expert observers and is very time consuming. Automatic or semiautomatic methods that may accelerate this process are therefore desirable, provided accuracy is not compromised.

In this work, we focus on the problem of determination of tumor volumes in the liver. The liver is frequently the first organ to show signs of metastatic spread for a range of cancers, and chemotherapy is a common treatment for patients at this disease stage. Thus improvements in methods to assess response in liver disease are likely to be of clinical significance. Additionally, alternative treatments such as the use of tumor ablation with radiofrequency devices also have the potential to be quantified using tumor volumetric evaluation.¹⁶

While liver lesions are commonly assessed with contrast-enhanced CT, many patients also undergo combined PET/CT. In PET/CT, the PET information is used to determine the functional status of the tissue, while the CT is used for PET attenuation correction and localization. When CT is used for attenuation correction, there is some debate as to whether contrast should be employed as it can potentially impact the estimates of PET radiotracer concentration.¹⁷ If contrast is not used, liver lesions can be harder to visualize on CT but there would be an advantage to the patient in terms of radiation dose if the nonenhanced attenuation-correction CT could

be used to determine tumor volume. We explore this possibility in this work.

A variety of methods for semiautomated tumor segmentation have been suggested in literature. These include thresholding and edge detection,¹⁸ region-growing techniques¹⁹ including the watershed method^{20,21} and fuzzy c-mean clustering,²² active contours or snakes,²³ and gradient vector flow (GVF) snakes.²⁴ Additionally, initial studies have been done on performing automatic segmentation on CT scans to determine liver tumor volume.²⁵

Mancas and Gosselin^{26,27} proposed a semiautomatic method which is a hybrid of the watershed and GVF snake methods.²⁸ Previously, a modified version of this method has been applied in segmenting lymphomas.²⁹ In this approach, operator defined internal and external markers are used to reduce region fragmentation. The watersheds are computed on the GVF field rather than simply the image gradient in order to suppress false edges and noise artifacts. Mancas and Gosselin²⁶ also added an iterative component allowing the delineation of “fuzzy” segmentation regions. Prior information is used as a guiding factor affecting segmentation. In the original implementation, iterative watershed segmentation (IWS) was applied on a slice-by-slice basis, requiring the setting of internal and external markers on each image slice. We refer to this as two-dimensional IWS (2D-IWS). However, the method may be generalized to operate on an image volume set so that, in principle, internal and external markers need only be set once per tumor, and spiculations or concavities oriented in the *z*-direction may be more easily accounted for. We refer to this method as three-dimensional IWS (3D-IWS).

The main objective of this paper is to compare the accuracy and feasibility of 2D-IWS and 3D-IWS in the task of determination of liver tumor volumes from CT images acquired both with and without contrast agents using the accuracy of physician observers as a reference. The effects of varying parameters (i.e., contrast, noise type, variance, and tumor size) on segmentation for simple spherical and realistic shape tumor models are described. The algorithms are applied to real CT images and a preliminary investigation of observer variability (intra and inter) in volume estimation is also performed for comparison.

II. CONSTRUCTION OF IWS

II.A. Watershed method

The watershed method (WM) is a region-growing segmentation technique. With the grayscale image being analogous to a topological surface, the region-growing method is similar to flooding catchment basins representing local minima from the image. As a result, ridge lines are formed around each basin, separating the regions from one another. Frequently, WM is performed on the gradient map of the image than directly on the pixel values.

For images that have low noise, WM is particularly effective since it can detect blurred object edges better than the simple methods of thresholding and edge detection or a combination of both.²⁰ However, medical images often have high

FIG. 1. Profile view of image gradient using GVF external field. Markers A and B are seeding points for flooding with watershed W being the resultant line.

overall noise variance resulting in a high density of local minima, and this may lead to oversegmentation.²⁶ The standard way to overcome this problem (and the one implemented here) is to add seeding markers to delineate where to start WM flooding. This implementation falls into the “ordered queue” class of WM, as described by Beucher *et al.*³⁰

II.B. The gradient vector flow (GVF)

Xu and Prince³¹ developed the GVF active contour “snake” as a modified version of classical active contours that tries to overcome the two main disadvantages that plague classical active contours in terms of segmentation performance—that is, lack of capture range and difficulty with concave boundaries. Summarized, the GVF algorithm iteratively solves the generalized diffusion equations to produce the GVF transformation result for the entire image. The user then applies the snake as a deformable parametric curve that uses iterative energy minimization on the GVF transformed image to segment the selected object. In the hybrid

IWS approach, active contours are not used, but the watersheds are computed on the GVF field transformation of the image.

II.C. IWS methodology

The WM with markers is illustrated conceptually using a profile view of the catchment basins in Fig. 1. In this implementation, the operator picks a small set of points definitively outside the tumor. Lines joining these points are considered as the set of external marker points (A). The operator then picks a small set of points definitively inside the tumor and lines joining these points are considered as the set of internal marker points (B). The watershed method with markers is then applied, resulting in two segmentation regions, “inside” the tumor and “outside.” In the iterative context, this is known as “level 0” segmentation.

Figure 2 illustrates a transverse view of the image segmentation results on a CT slice containing a tumor. For the next iteration (“level 1”) the tumor boundary (indicated as watershed W) from level 0 is considered as a third set of markers, and the watershed method applied again using markers A and B, and watershed W. This results in three segmentation regions: Inside the tumor (I), a fuzzy region (F), and outside the tumor (all image area not part of I and F). Likewise, this is repeated for IWS level 2, where there are now five sets of markers with the two new WM lines added. There will now be five regions of inside, three fuzzy regions, and outside the tumor. Progressing toward the inner markers, these fuzzy regions have an increasing degree of membership of the tumor category.

A general rule is that the number of markers set is equal to the number of segmentation regions. More iterations may be performed but in this work the limit is set to level two. This limit is chosen for practical reasons. More iterations result in larger numbers of boundaries, and therefore more statistical uncertainty in the comparisons. Also, the process of selecting a boundary from many options becomes unwieldy, potentially reducing the time-saving benefits of the technique.

FIG. 2. The first image from the left shows the initial seeding marker sets A and B along the resultant level 0 watershed line W. The second and third images show all subsequent level 1 and 2 resultant watershed lines, respectively.

II.D. Implementation details

The implementation assessed here is a modification of the publicly available demonstration code of Mancas and Gosselin,²⁶ utilizing the previously written code for the WM (Ref. 32) and GVF (Ref. 33) algorithms. In their approach, prior information using basic anatomic delineation rules (i.e., tumors not existing in airways, outside the body, in bone) and quick tumor localization arising from functional imaging (FDG PET) is utilized for marker selection.³⁴ In this work, the *a priori* model was removed. In the GVF subroutine used in the demonstration, the user-set parameters of μ and the number of iterations are set to 0.2 and 10, respectively. μ is the regularization parameter in the energy minimization functional that defines the GVF field and operates as a smoothing term, while the number of iterations dictates how many loops are used to solve generalization diffusion equations needed to find the GVF field. For our implementation, these values were left to their default values from the demonstration code. Additionally, the dimensionality was extended to support the 3D-IWS model and image assessment region was cropped to a 140×140 pixels per image slice to reduce computational load.

III. MATERIALS AND METHODS

III.A. Patient data

Patient data were selected from routine clinical evaluations representative of the typical clinical population suffering from metastatic liver disease at UC Davis Medical Center. These scans were acquired using various clinical CT imaging scanners and from a single PET/CT scanner (Discovery ST, GE Medical Systems, Waukesha, WI). For patients not scanned on the PET/CT scanner, 94–186 ml of IV contrast agent was injected at a rate of 2 ml/s using a standard IV contrast bolus route. These CT images have a slice thickness of 5 mm for the contrast-enhanced CT scanned cases and 3.75 mm for non-contrast-enhanced PET-CT cases. Pixel spacing ranged from 0.545 to 0.976 mm. Tube voltage ranged from 120 to 140 kVp and tube current from 280 to 400 mA. All images are saved in digital imaging and communications in medicine (DICOM) format with pixel values in Hounsfield units (HU) and have a matrix size of 512×512 pixels. Overall, there were 13 liver lesions for CE-CT and 7 liver lesions for PET-CT chosen from 11 patients.

III.B. Observer studies

MATLAB 7.0 with Image Processing Toolbox 4.2 (Ref. 35) was used to develop the graphical user interface (GUI) that allows manual outlining of tumors per slice. These outlines can be loaded and saved; in addition, image sets can be loaded in DICOM format. There are built-in window and level controls that allow the operator to select conventional radiological windows and to make manual adjustments to

them. Images may be magnified for region drawing. The outlines on every slice can have individual vertex points moved around interactively, along with vertex addition and deletion. This GUI is fairly straightforward to use after receiving an initial demonstration. All outlines were performed on Windows PCs.

In performing observer studies, lesions of a range of sizes and contrasts were initially selected from the patient images by a physician (R.H.). Four physicians were assigned to outline visible tumors in patient data (R.H., M.C., S.S., and M.G.). Each physician performed outlining on the same tumor for three separate instances, with each outlining instance spaced apart by at least 7 days. Using the GUI tool, each tumor was individually outlined and saved.

In addition to providing the data for observer studies, these outlines were also used to generate internal and external marker sets for applying IWS to real patient data. For each tumor, the intersection and union of three observers' outlines for one particular instance were determined to give observer-agreed outline areas. Generation of the fourth observer's outlines was delayed, so in the interests of time only the first three observers' data were used for this purpose. Additionally, the outlining instance was rotated from first to third per tumor to prevent biasing based on instance. The inner marker box was formed by taking the bounding box that encapsulates the regions' intersection and scaling to 20% of its original size while being centered around the centroid (see Fig. 3). The external marker box was formed by taking the bounding box that encapsulates the regions' union and adding 3 pixels to the corners to add buffer space between the intersection boundary and outer marker box. Using the outline's union for the outer marker box prevented any boundary crossings with the observer outlines from the marker boxes.

III.C. Creation of software phantoms

To examine the impact of noise magnitude and quality, object size, and object shape on the segmentation methods, a series of three-dimensional software phantoms were created. Each one contained a single region modeling a lesion embedded in a uniform background with noise added. The various parameters for each software phantom are listed below with Fig. 4 providing slice samples.

III.C.1. Tumor volume

A series of spherical tumor models was generated with sizes governed by the range found from the manual outlines of tumors in the patient data (minimum volume: 200 mm^3 , maximum volume: $102\,400 \text{ mm}^3$, mean volume: $21\,664 \text{ mm}^3$). A range of nonspherical models were also generated from the outlines of regions drawn by the observers over real examples from patient data. From a set of 12 individual tumor outlines, 10 shapes (excluding the biggest and smallest shape) were chosen and resized to a specific volume bin (800, 1600, etc., to 102 400) for a total of 8

FIG. 3. One slice instance of manual tumor segmentation regions from three radiological operators and the overall resultant intersection and union areas indicated by the image titles. The outer marker is displayed as a surrounding box in the regions' union image while the inner marker is displayed as an inner box in the regions' intersection image.

volume bins. Each selected tumor was selected by concurrently cycling through each replication instance and physician to reduce observer and instance bias.

III.C.2. Tumor contrast

A range of contrasts were simulated, matching those of the tumors analyzed in the patient data. The contrast difference was calculated by subtracting the mean tumor ROI value from the mean background value of four separate ROIs (each 50 pixels in area) drawn over background tissue and visually selected to be typical and homogenous. The average contrast difference was 21.95 HU with a range spanning from 7.64 to 42.55 HU. It should be noted that in the population examined there is likelihood of selection bias toward tumors with high contrast due to their enhanced visibility.

III.C.3. Spatial resolution

Spatial blurring was performed on the images with a convolution kernel representative of CT image acquisition and reconstruction blurring. For this purpose, a representative modulation transfer function (MTF) curve was chosen from the literature.³⁶ This MTF was then converted to a point spread function (PSF) and fitted to a Gaussian to be used for blurring. The resultant PSF had a sigma of 4.37 mm.

FIG. 4. The top two images are slice snapshots of software phantoms with added Gaussian noise (left) and correlated noise (right). The CNR of both images are set to 3.5 with all other parameters set equal. The bottom two images are slice snapshots of two Gaussian noise software phantoms with contrast difference values equal to 22 HU (left) and 32 HU (right). All other parameters are set equal.

III.C.4. Noise

Noise was modeled either as a series of Gaussian noise replicates or as correlated noise (derived from reconstructed image data—see below) and added to the blurred images containing the tumor models. The magnitude of the Gaussian noise was chosen to cover the range of noise seen in the patient images, determined by computing the variance in uniform regions of the liver. Ten Gaussian replicates were generated for each noise level. For the patient images, the noise variance value was found by taking the tumor ROI (from a sample slice) and by finding the standard deviation of the ROI intensity distribution. From the clinical patient data sampling, the average tumor ROI standard deviation was found to be 17.75 HU with a range of 11.10–24.67 HU.

Correlated noise was derived from CT scans of an adult chest-sized uniform water phantom³⁷ performed on a 16-slice PET/CT scanner (Discovery ST, GE Medical Systems, Waukesha, WI). Pixel spacing was 0.75 mm and slice thickness was 3.75 mm. The tube voltage was set to 120 kVp and the noise magnitude was varied by setting the tube current from 275 to 400 with increments of 25 mA. Ten replicates were acquired for each tube current level. This model assumes that the tumor has a negligible effect on the noise. The top images in Fig. 4 show images generated using the two noise models.

III.D. Comparison of 2D and 3D implementations

III.D.1. Interpolation

Prior to segmentation by 3D-IWS, image data were rebinned into isotropic cubic voxels of dimension equal to the default pixel spacing using trilinear interpolation. No rebinning was performed on the data prior to 2D-IWS segmentation. In performing 3D segmentation, resampling anisotropic data into isotropic cubic voxel data is commonly done^{14,38} primarily in maintaining the key assumption in calculating the GVF external field that the image pixels or volumetric voxels are isotropic. If the slices comprising of a volume set are stacked together without cubic voxel interpolation, there is the appearance of discontinuities in the z -direction of the 3D GVF external field due to instantaneous changes between neighboring slices' voxels. Another reason for cubic voxel interpolation is to lessen the effects of segmentation “bleeding.” This happens when separate segmentation regions are able to bypass the user-set markers in the x - y plane by moving around them in the z -direction. Stacking the anisotropic voxel images on top of one another without rebinning can lead to a distortion of the distance metric in the z -direction, exacerbating this bleeding effect. In our implementation the markers themselves were stretched into bands to match the interpolated cubic voxels, further lessening the bleeding effect.

FIG. 5. Diagram of three possible resultant watershed lines found by taking IWS on a sample image that are most likely closest to the actual tumor boundary. The ground-truth outline is the dashed line with the initial inner and outer markers indicated by Inner Marker B and Outer Marker A, respectively. The first level 0 outline is indicated by W0 with level 1 and level 2 outlines indicated by W1_1 and W2_2, respectively. The three outlines are typically closest to the ground-truth outline.

III.D.2. Amount of marker information

For 2D-IWS, one internal and external marker set was used for each slice. For 3D-IWS, different marker configurations were assessed as follows.

- All slices rule (3D AS): One internal and external marker on each slice with two additional markers (at the top and bottom) to delineate extent in z (analogous to 2D).
- Third rule (3D TR): One internal and external marker on every third slice with two additional markers (at the top and bottom) to delineate extent in z .
- Central slice rule (3D CS): One internal and external marker on only the central slice (minimal user information).

III.E. Effect of marker placement on volume estimation

An experiment was performed on the CE-CT data to assess the dependence of the volume calculation on the marker placement. This was done by separately examining the two aspects of scaling and shifting in regard to the inner and outer marker boxes. Scaling was performed by either adding/subtracting more than the original 3 pixels (6 and 9 pixels) from the corners of the outer marker box or scaling the inner marker box to be more than the original 20% of the regions' intersection bounding box (30%). Shifting was done by moving either of the marker boxes in one of four diagonal directions two pixels from the centroid. After performing one of

these operations, percent difference error was calculated between the IWS volumes using the original marker setting rule and those resulting from the selected operation.

III.F. Comparison with observer data

For the observer studies, tumor volume estimates were summarized initially for each tumor by box plots, means, and standard deviations. Tumor volume was calculated by taking the sum of all drawn slice contours and multiplying it by the slice thickness. As expected, the raw volume data showed a skewed distribution with a long tail toward a few larger tumors. All subsequent analyses used log-transformed data,³⁹ and results were summarized in terms of percent difference. Repeated measure analysis of variance (ANOVA) was used to compare physicians for consistent difference in size estimates, after adjusting for the size of a given tumor, and to test whether these differences were consistent across tumors or might vary from tumor to tumor (physician-tumor interaction). The residuals from this mixed model were then examined to determine whether physicians showed a tendency to converge on later viewings by comparing the absolute value of the difference between first and second reading with the absolute value of the difference between second and third reading using a paired t-test. Studentized range tests were used to compare group means found to differ significantly in ANOVA. All analyses were carried out using SAS/STAT® software⁴⁰ and graphics produced in R.⁴¹

Individual segmentation result points were plotted on top of each tumor residual difference box plot range. The number of segmentation data points within the box plot minimum-maximum range is tabulated in a chart. A high number of data points within these ranges would be considered to strengthen the case for one or both versions of IWS being able to act as a viable tumor segmentation method.

IV. RESULTS

IV.A. IWS performance with spherical and realistic shape models

IWS was not able to robustly segment the two smallest volumes (200 and 400 mm³) in the phantom studies. Very large variances in the volume estimations were found which on visual inspection appeared to be due to the algorithm being unable to distinguish the true boundaries from noise. Results for these volumes were then excluded from further analysis.

Typically, the segmentation outline from level 0 tends to underestimate the ground-truth volume when following the GVF edge map transformation. So as IWS progresses with further levels of segmentation performed, the outermost volume outline for level 1 (called level 1_1 consisting of inner region plus fuzzy region) and the second outermost outline for level 2 (called level 2_2 consisting of inner region plus two fuzzy regions) tend to have smaller percent error values than the initial level 0 outline. This outlining order concept is illustrated in Fig. 5.

Repeated measure ANOVA was used to examine the effects of the four segmentation methods (2D-IWS and all three marker configurations of 3D-IWS) and the three previously described level iterations in the presence of varying levels of contrast (experiment 1), Gaussian noise (experiment 2), correlated noise (experiment 3), and size and shape of lesion (experiment 4). Bonferroni-corrected 95% confidence intervals were created to compare level iterations 0, 1_1, and 2_2 within each experiment. The main goals were to determine the existence of differences between the three possible level iterations; a secondary goal is to determine which level has the smallest error. Several transformations of the data were considered to ensure that model assumptions were met. The results are summarized in Table I and Fig. 6. This analysis was performed in R 2.4.1. A significance level of 0.05 was used.

IV.B. Observer data trends: Inter- and intrarater variability

ANOVA showed that all physicians read consistent differences in size across tumors; the percent difference between estimated size of two tumors did not differ significantly across physicians ($P=0.75$). The physicians did differ systematically, however, with one physician (M.C.) consistently reading tumors about 12% smaller than the other three physicians (a 0.11 difference in means on the natural log scale). This seems consistent with the fact that M.C. was the only nonradiologist operator with the least diagnostic radiology training (nuclear medicine resident) among the group.

Physicians developed greater internal consistency after repeated measurements of the same tumor, with the third reading being on average about 17% closer to the second reading than the second was to the first ($P<0.001$). This is indicative of a training effect and means that the precision of manual volume estimation is overestimated.

IV.C. Effect of marker placement on volume estimation

Table II shows the average absolute percentage variation in volume estimation as a result of marker placement variation in the CE-CT data. Generally, IWS volume estimation appears reasonably robust with respect to variations in the marker placement, with an average absolute variation (over all conditions) of 3.0%. When shifting the markers, we found less variation for tumors that were >1000 mm³ in volume, although there was one outlier where a variation of 32.6% was seen. When expanding the markers, the trend of increasing reliability with tumor size appears reversed, but again this is due to a single outlier in the larger tumor group, where a maximum variation of 38.4% was found. This, the largest variation seen in experiment, corresponds to approximately an 11% variation in a linear measurement (assuming spherical tumor geometry).

TABLE I. Results of repeated measure ANOVA comparing mean log absolute error for iterations and segmentation methods, adjusting for variations in image quality and in lesion shape and volume. All confidence intervals are Bonferroni adjusted. Effects of level are shown for reference values of the factor varied.

ANOVA table				Effects of level (estimate, 95% CI)	
Expt. No.: Factor varied	Parameter	F-value	p-value	1_1 vs 0	2_2 vs 0
Expt. 1: Contrast difference	Value	51.8	<0.0001		
	Method	12.01	<0.0001		
	Model	88.58	<0.0001		
	Level	469.25	<0.0001	0.46 (0.34,0.59)	-1.22 (-1.34,-1.09)
Expt. 2: Noise variance-Gaussian	Value	8.92	<0.0001		
	Method	5.04	0.0022		
	Level	495.68	<0.0001	0.49 (0.35,0.64)	-1.41 (-1.55,-1.27)
Expt. 3: Noise variance-Correlated	Value	1.25	0.2866		
	Method	16.35	<0.0001		
	Level	512.62	<0.0001	0.51 (0.36, 0.66)	-1.5 (-1.65,-1.35)
Expt. 4: Lesion shape and volume	Value	355.58	<0.0001		
	Shape	2.09	0.1483		
	Method	4.24	0.0055		
	Model	16.24	0.0001		
	Level	64.54	<0.0001	0.05 (-0.06,0.17)	-0.48 (-0.59,-0.36)

IV.D. Comparison of IWS with observers

Figure 7 shows the range of observer volume estimates for CE-CT and PET-CT data compared to the mean for each tumor, with the level 2_2 and 1_1 IWS results overlaid. The complete results for each level and IWS version are summarized in Table III.

An additional row in this table displays the “best result” values, where, for each tumor, the iteration level resulting in the volume estimate that is closest to the average volume per

tumor is selected. This simulates the possible practice of the operator picking the best outlining instance based on his/her outlining expertise.

In general, 2D-IWS provides more data points within the observer ranges than the third rule version of 3D-IWS, especially for the CE-CT data. Alternatively for the PET-CT data, there was only one instance (level 1_1) where more 2D-IWS results (versus 3D-IWS TR) are within the corresponding manual volume tumor range. Among each level for the

FIG. 6. Accuracy of segmentation methods and levels for different experimental conditions.

TABLE II. Results displaying absolute percent differences in IWS volumes after performing a shifting or scaling operation on one of the marker boxes compared to IWS volumes using the original marker placing rule. Each individual result is the average percent difference from the CE-CT tumor set. In each case one marker was modified while the other marker was held constant.

	Level 0 (average % change)	Level 1_1 (average % change)	Level 2_2 (average % change)	Maximum variation (%)
Scaling one marker box ($\text{vol} < 1000 \text{ mm}^3$)				
Outer marker expanded by ± 6 pixels	1.44	2.16	1.83	6.92
Outer marker expanded by ± 9 pixels	2.18	2.43	2.55	7.14
Inner marker expanded from 20% to 30%	0.72	0.61	1.56	7.37
Scaling one marker box ($\text{vol} > 1000 \text{ mm}^3$)				
Outer marker expanded by ± 6 pixels	2.39	3.99	6.33	21.76
Outer Marker expanded by ± 9 pixels	4.23	7.10	5.92	38.40
Inner marker expanded from 20% to 30%	1.66	0.63	1.11	5.25
Shifting marker box from centroid ($\text{vol} < 1000 \text{ mm}^3$)				
Inner marker shifted ($x \pm 2 \text{ pixel}, y \pm 2 \text{ pixel}$)	8.12	5.97	7.28	23.26
Outer Marker shifted ($x \pm 2 \text{ pixel}, y \pm 2 \text{ pixel}$)	3.60	5.47	4.96	16.46
Shifting marker box from centroid ($\text{vol} > 1000 \text{ mm}^3$)				
Inner marker shifted ($x \pm 2 \text{ pixel}, y \pm 2 \text{ pixel}$)	1.34	0.76	0.97	7.25
Outer Marker shifted ($x \pm 2 \text{ pixel}, y \pm 2 \text{ pixel}$)	0.95	1.76	2.36	32.64

CE-CT data, level 1_1 provided the most number of points within the ranges, while level 0 (conventional IWS) produced the lowest number of points. Level 2_2 provided the highest number of points within the ranges for the PET-CT data.

V. DISCUSSION

For the phantom studies, ANOVA showed that level iteration had a strong effect on precision across all experiments, accounting for a large fraction of the variability in accuracy across measurements. Level 1_1 estimates have a larger error than those at level 0 (although not always a significant difference as seen in experiment 4). Level 2_2 estimates are more precise than those at level 0.

Segmentation method also showed significant, consistent differences in accuracy in all experiments, although these accounted for substantially less variation. The absolute error between measurement and true volume also varied with the size of the true tumor, as expected (Table I, experiment 4). Contrast difference also affected the absolute error in measurement (experiment 1), but the effects of noise were not consistent. The amount of noise affected the absolute difference when the noise model was Gaussian but not when it was correlated (experiments 2 and 3). This is likely due to the smaller variance range in the correlated noise and is encouraging—it suggests that the segmentation method is reasonably robust with respect to noise, at the noise levels likely to be encountered in practice.

A key factor that was examined was the placement of the inner and outer marker boxes. Separately shifting and scaling the marker boxes resulted in some variation in the IWS volume estimates. On average, this uncertainty was less than 5%, but there were individual cases of significantly larger

variation. This indicates that in practice IWS volume estimation will not be entirely operator independent.

For the phantom studies, 3D methods were generally more accurate than 2D, while for the observer studies, 2D-IWS is for the most part more accurate. The PET-CT data show data trends where 3D-IWS has more points in range versus 2D-IWS for certain level iterations; however, this may be anomalous due to the small number of PET-CT tumors. Average error is larger for the PET/CT data—this is consistent with the absence of contrast enhancement.

A common trend that reveals itself throughout both the phantom and *in vivo* studies is that 3D-IWS outputs smaller volumes than 2D-IWS when comparing the same level iteration. The cause of this result is not entirely understood but may relate to the facts that (a) features in the GVF in one slice directly affect those in adjacent slices, (b) the 3D data

TABLE III. Contained number of individual IWS data points within the gold-standard ranges for each tumor.

	2D	3D TR	3D AS	3D CS
CE-CT IWS data points within range (out of 13)				
Level 0	8	1	2	2
Level 1_1	11	9	11	11
Level 2_2	11	7	9	8
Best results inclusive of all levels	13	10	12	12
PET-CT IWS data points within range (out of 7)				
Level 0	4	4	4	3
Level 1_1	3	2	4	2
Level 2_2	5	6	7	5
Best results inclusive of all levels	7	7	7	6

FIG. 7. Residual box plots with level 0, 1_1, and 2_2 IWS results for each tumor, divided into the two groups CE-CT and PET-CT. Upward and downward pointing triangles indicate level 0 2D-IWS and 3D-IWS TR results, respectively. Asterisks and stars indicate level 1_1 2D-IWS and 3D-IWS TR results, respectively, while the circles and squares indicate level 2_2 2D-IWS and 3D-IWS TR results, respectively.

was interpolated to give cubic voxels, and (c) the tumor shapes were for the most part convex.

The fact that the gold standards for volume estimation in the patient data were defined on image stacks that were not interpolated in the axial dimension may be considered a confounding factor for comparison of the 2D and 3D methods since the 3D method was assessed on interpolated data, while the 2D method was assessed on data identical to that upon which the gold standards were defined (thus giving the 2D method an “unfair” advantage). However, the primary purpose of this work is to compare the 2D and 3D methods in settings that model that which might occur in the clinic. It is hard to envisage a clinical scenario where radiologists would wish to estimate volumes on axially interpolated (and therefore blurred) images, so that assessing the 2D method in the uninterpolated context is probably more reflective of its accuracy in practice. Since the 3D method requires interpo-

lation to cubic voxels as a preprocessing step, this is perhaps more of a limitation of the 3D method itself than of the comparison of the two. In further work, not reported in detail here, we compared the performance the 2D and 3D methods on identical interpolated patient data and determined that while the difference in the accuracy of the 2D and 3D volume estimates was reduced, 2D remained more accurate, at least with respect to the gold standard defined by the observers on the noninterpolated images.

When taking into account the best results tabulation, the number of points within range for 2D-IWS increases from 11 (maximum for level 1_1 and level 2_2) to 13 out of 13 and from 5 (maximum for level 2_2) to 7 out of 7 for CE-CT and PET-CT, respectively. A similar increase occurs for 3D-IWS TR when going from the individual level iterations with the most points in range to the best results. This demonstrates a key difference between this and many other segmentation

FIG. 8. First middle-slice snapshot (a) (top six images) of a CE-CT tumor outlined by 2D-IWS where only the outermost segmentation iteration (level 2_1) matches the tumor outline similar to that selected by the physicians. The second middle-slice snapshot (b) (bottom six images) of another CE-CT tumor outlined by 2D-IWS has an alternative scenario where level outlines 2_2 and 1_1 are closest to the physicians' outlines.

methods in that the user has the ability to choose which outline iteration to pick by utilizing the concept of additional fuzzy regions native to IWS. This is illustrated in Fig. 8. The top image set shows a tumor with substructure, where only the outermost segmentation iteration (level 2_1) finds the tumor outline similar to that selected by the physicians. The

bottom image set shows an alternative scenario where levels 2_2 and 1_1 are closest to the physicians' outlines.

As noted above, the estimated observer variability is likely reduced compared to truth due to training effects. However, for the purpose of this study, such decreased variance simply provides a more conservative benchmark against

which to measure the accuracy of the IWS segmentation results.

While 2D-IWS appears to be more accurate than 3D-IWS when a subset of slices were used for selecting preliminary points, the benefits in terms of time savings have not been quantified here. The trend in radiology is currently toward generating more image slices with thinner slice thickness—thus any such benefits from 3D-IWS are likely to become greater over time.

This preliminary examination of IWS is limited to finding the absolute volumes of single tumors. In anticancer drug-therapy trials, the key benchmark being measured is the *change* in tumor size from baseline to post-treatment scans. In this context, errors in outlines could easily be correlated between imaging sessions, potentially leading to substantial reductions in the variability in size change estimates between observers and/or the semiautomatic methods. Further work should involve comparison of the size change estimates between IWS and observers in clinical trials and correlation of the results with patient outcomes.

VI. CONCLUSIONS

This paper has demonstrated the potential application of the semiautomatic segmentation algorithms of IWS using phantom and *in vivo* studies. These initial results suggest that computer-aided segmentation can at least equal the accuracy of manual segmentation volume determination while potentially saving overall operator involvement time. Among the observer studies, it appears that 2D-IWS overall produces slightly more accurate results than 3D-IWS. Further studies should be performed to test these segmentation methods in determining tumor size change during drug-therapy trials.

ACKNOWLEDGMENTS

This work was supported by NIH under Grant No. 5T32EB003827, by the U.C. Davis Health System Research Award Program grant, "Tumor Metrics in Cancer Therapeutics," and by the NCI IRAT supplemental award P30-CA093373-04S2. Statistical analysis was supported by Grant No. UL1 RR024146 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH. Information on re-engineering the Clinical Research Enterprise can be obtained from <http://nihroadmap.nih.gov/clinicalresearch/overview-translational.asp>.

^{a)}Electronic mail: sray@ucdavis.edu

¹S. Saini, "Radiologic measurement of tumor size in clinical trials: Past, present, and future," *AJR, Am. J. Roentgenol.* **176**, 333–334 (2001).

²K. D. Hopper, K. Singapuri, and A. Finkel, "Body CT and oncologic imaging," *Radiology* **215**, 27–40 (2000).

³P. Therasse *et al.*, "New guidelines to evaluate the response to treatment in solid tumors: European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada," *J. Natl. Cancer Inst.* **92**, 205–216 (2000).

⁴S. Green and G. R. Weiss, "Southwest Oncology Group standard response criteria, endpoint definitions and toxicity criteria," *Invest New Drugs* **10**,

239–253 (1992).

⁵G. D. Shah, S. Kesari, R. Xu, T. T. Batchelor, A. M. O'Neill, F. H. Hochberg, B. Levy, J. Bradshaw, and P. Y. Wen, "Comparison of linear and volumetric criteria in assessing tumor response in adult high-grade gliomas," *J. Neuro-Oncol.* **8**, 38–46 (2006).

⁶C. G. Moertel and J. A. Hanley, "Effect of measuring error on results of therapeutic trials in advanced cancer," *Cancer* **38**, 388–394 (1976).

⁷E. C. Kohn and L. A. Liotta, "Molecular insights into cancer invasion: Strategies for prevention and intervention," *Cancer Res.* **55**, 1856–1862 (1995).

⁸H. L. Kindler, "Moving beyond chemotherapy: Novel cytostatic agents for malignant mesothelioma," *Lung Cancer* **45**, S125–S127 (2004).

⁹A. Sgambato, A. Camerini, B. Faraglia, R. Ardito, G. Bianchino, D. Spada, A. Boninsegna, V. Valentini, and A. Cittadini, "Targeted inhibition of the epidermal growth factor receptor-tyrosine kinase by ZD1839 ('Iressa') induces cell-cycle arrest and inhibits proliferation in prostate cancer cells," *J. Cell Physiol.* **201**, 97–105 (2004).

¹⁰C. H. Holdsworth, R. D. Badawi, J. B. Manola, M. F. Kijewski, D. A. Israel, G. D. Demetri, and A. D. Van den Abbeele, "CT and PET: Early prognostic indicators of response to imatinib mesylate in patients with gastrointestinal stromal tumor," *AJR, Am. J. Roentgenol.* **189**, W324–W330 (2007).

¹¹K. Marten, F. Auer, S. Schmidt, G. Kohl, E. J. Rummeny, and C. Engelke, "Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria," *Eur. Radiol.* **16**, 781–790 (2006).

¹²A. G. Sorensen *et al.*, "Comparison of diameter and perimeter methods for tumor volume calculation," *J. Clin. Oncol.* **19**, 551–557 (2001).

¹³B. Zhao, L. H. Schwartz, C. S. Moskowitz, L. Wang, M. S. Ginsberg, C. A. Cooper, L. Jiang, and J. P. Kalaigian, "Pulmonary metastases: Effect of CT section thickness on measurement—initial experience," *Radiology* **234**, 934–939 (2005).

¹⁴B. S. Zhao, A. P. Reeves, D. F. 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).

¹⁵S. Rohde, B. Turowski, J. Berkefeld, and A. F. Kovacs, "CT-based evaluation of tumor volume after intra-arterial chemotherapy of locally advanced carcinoma of the oral cavity: Comparison with clinical remission rates," *Cardiovasc. Intervent. Radiol.* **30**, 85–91 (2007).

¹⁶D. L. Stippel, H. G. Brochhagen, M. Arenja, J. Hunkemoller, A. H. Holscher, and K. T. Beckurts, "Variability of size and shape of necrosis induced by radiofrequency ablation in human livers: A volumetric evaluation," *Ann. Surg. Oncol.* **11**, 420–425 (2004).

¹⁷G. Antoch, L. S. Freudenberg, T. Egelhof, J. Stattauss, W. Jentzen, J. F. Debatin, and A. Bockisch, "Focal tracer uptake: A potential artifact in contrast-enhanced dual-modality PET/CT scans," *J. Nucl. Med.* **43**, 1339–1342 (2002).

¹⁸S.-J. Park, K.-S. Seo, and J.-A. Park, "Automatic hepatic tumor segmentation using statistical optimal threshold," *Computational Science & ICCS 2005* (Springer, New York, 2005), Vol. 3514, pp. 934–940; P. Gibbs, D. L. Buckley, S. J. Blackband, and A. Horsman, "Tumour volume determination from MR images by morphological segmentation," *Phys. Med. Biol.* **41**, 2437–2446 (1996).

¹⁹B. Zhao, L. H. Schwartz, L. Jiang, J. Colville, C. Moskowitz, L. Wang, R. Leftowitz, F. Liu, and J. Kalaigian, "Shape-constraint region growing for delineation of hepatic metastases on contrast-enhanced computed tomograph scans," *Invest. Radiol.* **41**, 753–762 (2006).

²⁰M. M. Letteboer, O. F. Olsen, E. B. Dam, P. W. Willems, M. A. Viergever, and W. J. Niessen, "Segmentation of tumors in magnetic resonance brain images using an interactive multiscale watershed algorithm," *Acad. Radiol.* **11**, 1125–1138 (2004).

²¹E. Bellon, M. Feron, F. Maes, L. V. Hoe, D. Delaere, F. Haven, S. Sunaert, A. L. Baert, G. Marchal, and P. Suetens, "Evaluation of manual vs semi-automated delineation of liver lesions on CT images," *Eur. Radiol.* **7**, 432–438 (1997).

²²W. Chen, M. L. Giger, and U. Bick, "A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images," *Acad. Radiol.* **13**, 63–72 (2006); P. J. Yim, A. V. Vora, D. Raghavan, R. Prasad, M. McAuliffe, P. Ohman-Strickland, and J. L. Noshier, "Volumetric analysis of liver metastases in computed tomography with the fuzzy C-means algorithm," *J. Comput. Assist. Tomogr.* **30**, 212–220 (2006).

- ²³M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.* **1**, 321–331 (1988).
- ²⁴C. Xu and J. L. Prince, "Gradient vector flow deformable models" *Handbook of Medical Imaging* (Academic, New York, 2000), pp. 159–169.
- ²⁵L. Massotier and S. Casciaro, "A new fully automatic and robust algorithm for fast segmentation of liver tissue and tumors from CT scans," *Eur. Radiol.* **18**, 1659–1665 (2008).
- ²⁶M. Mancas and B. Gosselin, "Fuzzy tumor segmentation based on iterative watersheds," Proceedings of the 14th ProRISC Workshop on Circuits, Systems and Signal Processing (ProRISC 2003), Veldhoven, The Netherlands, 2003 (unpublished).
- ²⁷M. Mancas, "IWS Demo" (unpublished), <http://tcts.fpms.ac.be/rdf/demos/mateidemo/mateidemo.htm>.
- ²⁸P. J. Yim and D. J. Foran, Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems, 2003 (unpublished).
- ²⁹J. Yan, B. Zhao, L. Wang, A. Zelenetz, and L. H. Schwartz, "Marker-controlled watershed for lymphoma segmentation in sequential CT images," *Med. Phys.* **33**, 2452 (2006).
- ³⁰S. Beucher and F. Meyer, "The morphological approach to segmentation: The watershed transformation" *Mathematical Morphology in Image Processing*, edited by E. R. Dougherty (Dekker, New York, 1993), pp. 433–481.
- ³¹C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Trans. Image Process.* **7**, 359–369 (1998).
- ³²K. Forbes, "Some simple image processing tools for MATLAB" (unpublished), <http://www.dip.ee.uct.ac.za/~kforbes/KFtools/KFtools.html>.
- ³³C. Xu and J. L. Prince, "Active contours deformable models, and gradient vector flow" (unpublished), <http://iacl.ece.jhu.edu/projects/gvf>.
- ³⁴M. Mancas and B. Gosselin, "Towards an automatic tumor segmentation using iterative watersheds," Proceedings of the Medical Imaging Conference of the International Society for Optical Imaging (SPIE Medical Imaging 2004), San Diego, California, 2004 (unpublished), pp. 1598–1608.
- ³⁵MATLAB, the Mathworks, Natick, MA.
- ³⁶J. M. Boone, "Determination of the presampled MTF in computed tomography," *Med. Phys.* **28**, 356–360 (2001).
- ³⁷NEMA Standards Publication, No. NU 2-2001, Performance Measurements of Positron Emission Tomographs, Rosslyn, Virginia, 2001.
- ³⁸W. J. Kostis, A. P. Reeves, D. F. Yankelevitz, and C. I. Henschke, "Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images," *IEEE Trans. Med. Imaging* **22**, 1259–1274 (2003); M. Joliot and B. M. Mazoyer, "Three-dimensional segmentation and interpolation of magnetic resonance brain images," *ibid.* **12**, 269–277 (1993).
- ³⁹O. N. Keene, "The log transformation is special," *Stat. Med.* **14**, 811–819 (1995).
- ⁴⁰SAS/STAT®, Version 9.0, SAS Institute, Inc., Cary, NC, 2004.
- ⁴¹R Development Core Team, R, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2006.