Quantifying the margin sharpness of lesions on radiological images for content-based image retrieval

Jiajing Xu^{a)}

Department of Electrical Engineering, Stanford University, Stanford, California 94305

Sandy Napel

Department of Radiology, Stanford University, Stanford, California 94305

Hayit Greenspan^{b)}

Department of Biomedical Engineering, Tel Aviv University, Tel Aviv 69978, Israel

Christopher F. Beaulieu

Department of Radiology, Stanford University, Stanford, California 94305

Neeraj Agrawal

Department of Computer Science, Stanford University, Stanford, California 94305

Daniel Rubin

Department of Radiology, Stanford University, Stanford, California 94305

(Received 4 January 2012; revised 5 June 2012; accepted for publication 12 July 2012; published 16 August 2012)

Purpose: To develop a method to quantify the margin sharpness of lesions on CT and to evaluate it in simulations and CT scans of liver and lung lesions.

Methods: The authors computed two attributes of margin sharpness: the intensity difference between a lesion and its surroundings, and the sharpness of the intensity transition across the lesion boundary. These two attributes were extracted from sigmoid curves fitted along lines automatically drawn orthogonal to the lesion margin. The authors then represented the margin characteristics for each lesion by a feature vector containing histograms of these parameters. The authors created 100 simulated CT scans of lesions over a range of intensity difference and margin sharpness, and used the concordance correlation between the known parameter and the corresponding computed feature as a measure of performance. The authors also evaluated their method in 79 liver lesions (44 patients: 23 M, 21 F, mean age 61) and 58 lung nodules (57 patients: 24 M, 33 F, mean age 66). The methodology presented takes into consideration the boundary of the liver and lung during feature extraction in clinical images to ensure that the margin feature do not get contaminated by anatomy other than the normal organ surrounding the lesions. For evaluation in these clinical images, the authors created subjective independent reference standards for pairwise margin sharpness similarity in the liver and lung cohorts, and compared rank orderings of similarity used using our sharpness feature to that expected from the reference standards using mean normalized discounted cumulative gain (NDCG) over all query images. In addition, the authors compared their proposed feature with two existing techniques for lesion margin characterization using the simulated and clinical datasets. The authors also evaluated the robustness of their features against variations in delineation of the lesion margin by simulating five types of deformations of the lesion margin. Equivalence across deformations was assessed using Schuirmann's paired two one-sided tests.

Results: In simulated images, the concordance correlation between measured gradient and actual gradient was 0.994. The mean (s.d.) and standard deviation NDCG score for the retrieval of K images, K = 5, 10, and 15, were 84% (8%), 85% (7%), and 85% (7%) for CT images containing liver lesions, and 82% (7%), 84% (6%), and 85% (4%) for CT images containing lung nodules, respectively. The authors' proposed method outperformed the two existing margin characterization methods in average NDCG scores over all K, by 1.5% and 3% in datasets containing liver lesion, and 4.5% and 5% in datasets containing lung nodules. Equivalence testing showed that the authors' feature is more robust across all margin deformations ($p < 0.05$) than the two existing methods for margin sharpness characterization in both simulated and clinical datasets.

Conclusions: The authors have described a new image feature to quantify the margin sharpness of lesions. It has strong correlation with known margin sharpness in simulated images and in clinical CT images containing liver lesions and lung nodules. This image feature has excellent performance for retrieving images with similar margin characteristics, suggesting potential utility, in conjunction

with other lesion features, for content-based image retrieval applications. © 2012 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4739507>]

Key words: feature extraction, image retrieval, computed tomography (CT), lung nodules, liver lesions, margin sharpness

I. INTRODUCTION

A crucial challenge for radiology is maintaining high interpretation accuracy in the face of increasing imaging workload and limited time to review and interpret the images for each patient. Variation in interpretation accuracy among radiologists is a recognized challenge,^{1–3} and systems are being developed to help radiologists improve their interpretations.⁴ Content-based image retrieval (CBIR) could provide decision support to radiologists by allowing them to find images from a database that are similar in terms of shared imaging features to the images they are interpreting.

CBIR systems use text associated with images and image processing techniques to derive quantitative characteristics of images (e.g., pixel statistics, spatial frequency content), followed by application of similarity metrics to search image libraries for similar images.^{5,6} Outside of radiology, numerous CBIR methods have been developed for image retrieval and automatic image annotation.⁵ Most CBIR methods characterize entire images with a set of numerical features.^{7,8} Development of CBIR methods in the radiological domain has focused on retrieving images from medical collections.^{9,10} Recently, CBIR has been applied to localized image regions to retrieve images of similar-appearing lesions.^{11–16}

The performance of CBIR systems depends on extracting features that encapsulate the distinguishing characteristics of the image contents. In radiological images, texture features such as histogram features and wavelets have been commonly used,^{13,17–20} and they can capture features that may not even be visually apparent to radiologists.²¹ Beyond texture features, there are other features that can be very helpful in characterizing lesions. In particular, the margin characteristics of a lesion are known to be useful to radiologists for discriminating many types of lesions. Our goal in this paper is to develop a method to quantify the sharpness of the margin of lesions—a “margin sharpness feature”—and to evaluate its performance

for retrieval of CT images containing similar-appearing liver lesions and lung nodules. Prior work has been done in developing image features of lesion sharpness in breast, and we thus also compare our technique with margin features that has been proposed for assessing breast MRI lesions from the most cited²² and most recent works.²³

This paper is organized as follows: Section II describes our methods, including the margin sharpness feature, our test images, and system design for using our margin sharpness feature, Sec. III describes our results, and Sec. IV highlights key contributions and future work.

II. METHODS

II.A. System architecture and overview

Figure 1 shows the system design for developing and testing the proposed margin sharpness feature, comprised of three pipelines: (a) Database Building, (b) Query/Retrieval, and (c) Scoring for Evaluation. In the Database Building pipeline, we constructed a database containing CT images of circumscribed lesions in the liver and lung in which the lesions were circumscribed by hand. Given this annotated database, we submitted query images to the Query/Retrieval pipeline, in which margin features of each lesion were computed, and compared to those of the existing images in the database, producing a ranked list of images in the database in decreasing order of similarity (in terms of margin characteristics) to the query image. We used the Scoring for Evaluation pipeline to compare the resulting ranked list for each query image to an independent subjective standard for image margin similarity. Sections II.A.1–II.A.4 describe each of the components of our system and method for evaluation.

II.A.1. Margin sharpness feature

We use two attributes to define the sharpness of the margin of a lesion: (1) intensity difference: the difference between intensities of the organ tissue surrounding the lesion and the tissue inside of the lesion (“lesion substance”), and (2) margin blur: the abruptness of the transition in intensity from the lesion substance to the normal organ tissue surrounding the lesion. A sharper margin will have a more abrupt transition and may have a higher difference of intensities outside and inside the lesion, whereas a blurred margin will have a smoother transition and may have a smaller intensity difference (the intensity difference may be complicated by the contrast bolus—see Sec IV). Our feature to characterize the sharpness of the margin captures these two attributes. As described in Secs. II.A.2–II.A.4, each input image consists of a region of interest (ROI) circumscribing the lesion, which in our case was drawn

FIG. 1. System design.

FIG. 2. (a) Cropped CT image containing a lung nodule with a normal line and ROI drawn. (b) Pixel intensity values along the normal line (solid line) and fitted sigmoid (dashed line) showing the scale and window parameters.

by radiologists, but could also be generated automatically by a CAD system or automated segmentation algorithm. To generate the margin sharpness feature, we first represent the ROI by a piecewise cubic polynomial curve (interpolated from 10–20 control points drawn by a radiologist). We next draw normal line segments of length T pixels across the boundary of the lesion at fixed intervals around its circumference (the fixed interval is the larger value of 1 pixel and the length of the boundary in pixels divided by 200). Intensity values are then recorded along these segments using bilinear interpolation. We next fit a sigmoid function (Fig. 2) to these values using a robust weighted nonlinear regression.²⁴ For each line I , the problem can be formulated as

$$\arg \min_{S, W, x_0, I_0} \sum_x \left[I(x) - I_0 - \frac{S}{1 + e^{-\frac{x-x_0}{W}}} \right]^2,$$

where x is the distance along the normal, x_0 is the intersection of the boundary point with the normal, $I(x)$ is the intensity along the normal at x , and I_0 is the intensity offset.

Two parameters, scale (S) and window (W) from the sigmoid are then used to characterize each line segment I . The scale measures the difference in intensities outside and inside the lesion, and the window characterizes the margin blur by measuring the transition from the lesion to surrounding normal organ tissue at the boundary. We set the range of W to be

$-2*T$ to $2*T$ and the range of S to be -1000 HU to 1000 HU. If fitted parameters are out of the range, we do not include the result into the feature vector.

II.A.2. Margin sharpness feature implementation

The margin sharpness of each lesion is represented as a feature vector composed of [two] 30-bin histograms of the scale and window parameters obtained from each normal. We remove outliers by discarding values below the 5th percentile and above the 95th percentile for each parameter before constructing the histogram. Each histogram is normalized to have unit area. The feature vector is the concatenation of the two histograms.

It should be noted that in creating this computational representation of the lesion boundary, it is assumed that each normal line segment includes both lesion substance and normal organ tissue surrounding the lesion. However, for lesions at the edge of an organ, a portion of the circumference of the lesion will be adjacent to some structure other than the normal surrounding organ. Figure 3(a) shows such an example. Due to the possible changes in intensity between the organ and its surrounding tissue, the line segments drawn on the border of the organ do not have enough information to characterize the margin of the lesion. Including the scale and window parameters obtained from the line segments from this portion of the

FIG. 3. (a) Cropped lung CT image displayed at standard lung window-level setting, with ROI outlined (in dark gray) by radiologist. (b) Automatically segmented lung region (in gray) showing filling defect caused by inadequate segmentation near the lesion and the ROI outline (in white). (c) Final lung segmentation mask (the union of the automatically segmented lung mask and the lesion mask). (d) Cropped lung CT image with line segments (in white) that are completely inside the lung mask.

FIG. 4. (a) Cropped lung CT image containing a lung nodule in the center, displayed at standard lung window-level setting. (b) Same as (a), with a 3-pixel-thick shell marking where Gilhuijs's margin measurement is performed. (c) Same as (a), with two 1-pixel-thick shells marking where Levman's margin measurement is performed.

lesion circumference would introduce errors into the margin feature. We tackled this limitation by omitting the line segments drawn near or on the organ border.

For purposes of this paper, we obtained the liver boundaries manually and lung boundaries automatically. For each CT image containing liver lesions in our dataset, a radiologist traced the liver border. For each CT image containing lung nodules, we applied optimal gray-level thresholding to obtain a lung region mask from the CT images,²⁵ where voxels with value of 1 correspond to the voxels that belong to the lung and voxels with value of 0 correspond to the voxels that do not. As shown in Fig. 3(b), simple fully automated lung segmentation algorithms often perform poorly when severe lung disease is present in the patient.²⁶

In our work, since radiologists have already identified these lesions, satisfactory lung region segmentation in the region near the lesion was obtained by taking the union of the automatically segmented lung region mask and the lesion mask (where voxels with value of 1 represent the voxels that belong to the lesion) and filling any holes. Figure 3(c) shows the final lung segmentation mask. Once we obtain the organ masks (manually or automatically), we omit fitting a sigmoid to those line segments that are not completely contained inside the organ mask, illustrated in Fig. 3(d).

II.A.3. Prior work on characterizing edge features of lesions

Prior work on developing features to characterize lesion sharpness has been conducted, primarily in breast lesions. In breast MRI literature, the most commonly used mathematical methods for characterizing margin of lesions were first presented in Ref. 22. The method examines the three-voxel shell that marks the outer contour of the lesion. As shown in Fig. 4(b), Gilhuijs's two margin measurements, margin gradient and variance of margin gradient, are based on the average and the variance of the spatial gradient of the subtracted enhancement data from the light gray shell. The two features are

given by

$$M_{\text{Gilhuijs,mean}} = \max_{i=0,\dots,M-1} \left\{ \frac{\text{mean}_r \|\nabla I_m(r)\|}{\text{mean}_r I_m(r)} \right\},$$

$$M_{\text{Gilhuijs,var}} = \frac{\text{var}_r \|\nabla I_m(r)\|}{[\text{mean}_r I_m(r)]^2},$$

where $I_m(r)$ is the signal intensity of a voxel and the range of vector r in $I_m(\cdot)$ is limited to the three-voxel shell centered on the surface of the lesion.

Levman²³ recently proposed a similar method to use the voxels that are close to the edge of the lesion (but still in the lesion) and those that immediately neighbor the lesion (but are outside the lesion). This is demonstrated in Fig. 4(c), where the 1-pixel rings are colored dark gray and light gray for outside and inside the lesion, respectively. The margin measurement is defined as

$$M_{\text{Levman}} = \frac{I(r_i) - I(N(r_i))}{d},$$

where r_i represents voxels on the inner ring, $I(r_i)$ is the signal intensity of a voxel located at position r_i , $N(r_i)$ is the six-connect three-dimensional neighborhood operator, which provides a set of voxel positions that neighbor r_i but are outside the lesion ROI, and d is the normalization term.

To compare these existing edge feature methods to ours, we incorporated the organ masks in our image database as we did in Sec. II.A.2. In Gilhuijs's method, we excluded the segments of the three-voxel shell that cover any voxel from the border of the organ. In Levman's method, we excluded the voxel from the inner ring when its closest neighbor pixel is on the edge of the organ.

II.A.4. Robustness of the features

To evaluate how our feature might perform given likely inter-reader variation in circumscribing lesion margins, we investigated the robustness of our feature against several variations in the lesion margin. We created five types of deformations of the radiologist-drawn margin for each lesion as

FIG. 5. Examples showing the modified lesion contour (light gray) with respect to the original lesion contour (dark gray) under the following operation: (a) dilation (b) erosion (c) translation (d) rotation (e) random mutation. Circles are controls points of piecewise cubic splines.

follows: (1) dilation, (2) erosion, (3) translation, (4) rotation, and (5) random warp, as illustrated in Fig. 5. For (1) and (2), we dilated/eroded the lesion margin by $0.1 \cdot R_{\max}$, where $R_{\max}$ is the maximum diameter of the lesion. For (3), we shifted the lesion margin by $0.1 \cdot R_{\max}$ in a random direction. For (4), we rotated the lesion margin by 10° clockwise. For (5), we moved a random 30% of the control points that were used to generate the cubic spline of the lesion margin in a random direction by as much as $0.1 \cdot R_{\max}$. We chose these types and magnitudes of deformations to cover the space of expected tracing when all readers would agree the edge is roughly in the same place, but did not trace it exactly so. We then computed our margin sharpness feature for the images with deformed lesion boundaries, and compared the results with the manually traced lesion margin.

II.B. Image datasets

We used two types of images for testing the margin sharpness feature. Simulated images were generated to evaluate the correctness and robustness of the feature. We also built two datasets of clinical images containing lesions in two different body organs (liver lesions and lung nodules) to evaluate our feature.

II.B.1. Simulated images

The margin sharpness feature describes two aspects of the margin of a lesion: the margin blur as well as the intensity difference across it. To assess the ability of our feature to accurately represent these parameters, we ran experiments on 81 simulated circular lesions with known but varying margin sharpness. Each simulated lesion was characterized by two free parameters: (1) the difference between the average inten-

sity of the lesion substance and the surrounding organ (which for purposes of this experiment we assumed is liver), and (2) the amount of Gaussian blur applied to decrease the sharpness of the margin. To generate the image, a circular patch of darker intensity (lesion) was placed on a rectangular patch of lighter intensity (liver) with the appropriate Gaussian blur. The average intensity of the liver was set at 140 HU as is typical in portal venous liver CT scans. The difference in intensity between the lesion and the liver, and the degree of blur were varied independently with 9 levels each, thus generating 81 images. The blur levels corresponded to the size of the Gaussian kernel with standard deviations of 1–5 pixels. The intensity difference levels corresponded to a difference in intensity between the outside and inside of the lesion from 40 to 120 HU in increments of 10 (Fig. 6).

To make our simulated images realistic, we added correlated noise; this was produced using a water phantom.²⁷ Four water phantoms were scanned and a 5×5 autocorrelation matrix representing the noise correlation in CT scans was generated from each. Autocorrelation matrices from each image were then averaged to produce a single average autocorrelation matrix. Initially, Gaussian noise was added to the image. The noisy image was then convolved with the autocorrelation matrix to introduce correlated noise. We set the noise level to be close to that observed in abdominal CT scans (standard deviation is 10 HU).

II.B.2. Clinical images

Under IRB approval for retrospective analysis of deidentified patient images, we selected 79 portal venous phase CT images containing liver lesions and 58 CT images containing lung nodules from our hospital PACS system. The liver images were acquired on Siemens Sensation 64 (number of

FIG. 6. Simulated lesions with increasing margin blur from left to right and decreasing intensity difference across the margin from top to bottom.

cases, $n = 21$), Siemens Sensation 16 ($n = 7$), Siemens Volume Zoom ($n = 8$), GE Medical Systems LightSpeed 16 ($n = 19$), GE Medical Systems LightSpeed Qx/i ($n = 17$), and GE Medical Systems LightSpeed Ultra ($n = 7$) with mean current of 343 mA (range, 235–519 mA) and voltage of 120 kVp and slice thickness of 5.0 mm; the lung images were acquired on GE Medical Systems Discovery CT750 HD CT scanner ($n = 35$) and Siemens Sensation 64 ($n = 23$) with mean current of 340 mA (range 159–472 mA) and voltage of 120 kVp, and slice thickness of 1.25 mm. The 79 liver lesions (25 cysts, 24 metastases, 14 hemangiomas, three abscesses, one fat deposition, one laceration, five focal nodular hyperplasia, and six hepatocellular carcinomas) were derived from 44 patients (23 male and 21 female), mean age 61 yr old (range, 26–92 yr) and the 58 lung nodules were derived from 57 patients (24 males and 33 females), mean age 66 yr old (range, 32–84 yr). A radiologist circumscribed each lesion on the most representative cross section, generating an irregular ROI. For each liver CT image, the radiologist also delineated the liver boundary around the lesion.

II.C. Evaluation

II.C.1. Evaluating margin sharpness features in simulated images

Since we compare the performance of our margin sharpness feature to the two existing methods (Sec. II.A.3), and these measure the gradient along the lesion boundary, we computed the actual gradient along the lesion boundary in each of our simulated images. Although our feature does not measure the gradient directly, we derived the gradient from our feature by taking the ratio of the intensity difference pa-

rameter to the window parameter. We determined agreement between the gradient computed by our feature in this way and the actual gradient using the concordance correlation²⁸ and Bland–Altman analysis.²⁹ For the existing margin sharpness features (Sec. II.A.3), we performed the same evaluation on $M_{\text{Gilhuijs, mean}}$ and M_{Levman} vs the actual gradient in the simulated images.

For each of the three methods, to investigate the robustness of the derived features to the given location of the lesion boundary, we the concordance correlation of the features extracted using the original boundary to the ones extracted using modified deformed boundaries as discussed in Sec. II.A.4.

II.C.2. Evaluating margin sharpness features in clinical images

II.C.2.a. Generation of reference standard for clinical images. Five readers (three board certified, fellowship-trained radiologists with 20-yr, 19-yr, and 5-yr experience of abdominal imaging and two senior researchers in the medical imaging field) viewed each liver lesion image and five readers (four board certified, fellowship-trained radiologists with 20-yr, 19-yr, 15-yr, and 5-yr experience of abdominal imaging and one senior researcher in the medical imaging field) viewed each lung nodule image. The ROIs used in this setting were also used later for computing the edge feature for all methods. Each reader rated the “lesion margin definition” (how clearly or poorly defined the margin was visualized) for each image on a scale of 1 to 5 (1 = poorly defined margin and 5 = clearly defined margin). They were instructed to consider only the definition of the margin of the lesion in making this rating. They did not consider the boundary shape, size, or location

within the organ, nor did they consider any clinical data that might have implied a specific lesion type. Given these scores, we defined the reference standard of similarity in margin characteristic between two images i and j as

$$S_{ij} = 5 - \frac{1}{5} \sum_{k=1}^5 (|R_{ik} - R_{jk}|),$$

where R_{ik} is the rating given by k th reader to i th image. The reference standard similarity scores are in the range of [1–5] with 1 being least similar. We defined the computed similarity of a pair of lesions as the inverse of the sum of differences between corresponding elements of the respective feature vectors that describe them.

We assessed interobserver variability in the reference standard for the clinical images using Fleiss' kappa test.³⁰ This test evaluates the consistency of raters using categorical ratings and returns a score from 0 to 1, with 0 denoting complete disagreement and 1 denoting complete agreement.

II.C.2.b. CBIR experiments. We performed a series of queries using one image at a time from our collection to rank order the remaining images (performing separate queries for the liver lesion collection and the lung lesion collection), and we used the normalized discounted cumulative gain (NDCG) measure³¹ to evaluate the ranked output for each query. Discounted cumulative gain (DCG) takes K as a parameter and assigns a score to each permutation of the top K entries in the ranked list based on the number of violations of a perfect ranking in the output. A maximum score is achieved when, for a given query image, no images of lower similarity score in the reference standard appear before higher scores in the ranked list of K images (e.g., the score decreases when an image with similarity 3 follows an image with similarity 2). DCG also weights higher positions more than the lower ones. Thus, any out-of-order images in the top few positions have a higher impact than out-of-order images in the last few positions. Normalization is carried out by dividing DCG by the maximum attainable DCG for the particular output to yield a NDCG score for the top K positions of a ranked output. The NDCG score varies from 0% to 100%, higher being better.

II.C.2.c. Statistical analysis on NDCG scores. We used the paired Wilcoxon test³² to test the null hypothesis that there is no significant difference in the NDCG(K) scores obtained from two different methods (i.e., our proposed and Gilhuijs's method, our proposed and Levman's method). We performed the Wilcoxon tests at $K = 5$ and $K = 10$, since radiologists are more interested in these top ranked images.

To investigate the robustness of the features to variations in how the margin is drawn, we applied boundary deformations to each query image, as described in Sec. II.A.4. For testing for robustness of NDCG scores across the deformed lesion outlines, we used Schuirmann's paired two one-sided equivalence tests (TOST) (Ref. 33) for determining equivalence of NDCG scores obtained using original outlines and deformed outlines. We tested the null hypothesis that the mean NDCG score obtained from one method differs at least 20% from the one obtained from the other method.

FIG. 7. Bland–Altman plot showing the mean difference in the measured gradient by the proposed method and the actual gradient (solid middle line) and 95% limits of agreement (dashed lines). Unit is HU/pixel.

II.D. Parameter sensitivity experiments

In Sec. II.A.2, we recorded pixel intensity values along the normal line segment of length T pixels. Based on empirical studies, we have selected T to be 10. In order to evaluate the sensitivity of the choice of T , we investigate the average NDCG scores among top 10 retrieved images with respect to T , as T varies from 6 to 15 with an interval of 1.

III. RESULTS

III.A. Experiments with simulated images

The Bland–Altman plot (Fig. 7) shows consistent measurement of the actual gradient by our method with a mean difference of 0.61 HU/pixel and 95% limits of agreement of 5.3 HU/pixel. The concordance correlation between the measured gradient and the actual gradient was 0.994 (95% confidence interval 0.991, 0.996).

Figures 8(a) and 7(b) show Bland–Altman analysis for Gilhuijs's and Levman's method. For Gilhuijs's method, the mean difference is -6.09 HU/pixel and 95% limit of agreement of 16.33 HU/pixel. For Levman's method, the mean difference is -10.44 HU/pixel and 95% limit of agreement of 14.37 HU/pixel. As summarized in Table I, the mean concordance correlation between Gilhuijs's measured gradient and the actual gradient was 0.900 and the mean concordance correlation between Levman's measured gradient and the actual gradient was 0.881.

Table II summarize the concordance correlation between the actual gradient and the measured gradient by (1) the proposed method, (2) Gilhuijs's method, and (3) Levman's

FIG. 8. (a) Bland–Altman plot showing the mean difference in the measured gradient by Gilhuijs’s method and the actual gradient (solid middle line) and 95% limits of agreement (dashed line). (b) Bland–Altman plot showing the mean difference in the measured gradient by Levman’s method and the actual gradient (solid middle line) and 95% limits of agreement (dashed line). Unit is HU/pixel.

method after the deformations to the lesion boundary discussed in Sec. II.A.4. Our feature predicts the gradient parameter quite accurately with very high concordance correlation (greater than 0.98 in all cases), in scenarios with and without a deformed lesion boundary. Gilhuijs’s and Levman’s measured gradient features correlated with the actual gradient parameter, in general, but these methods are not as robust as our method to the change in lesion boundary. The concordance correlation using Gilhuijs’s method and Levman’s method for scenarios with a deformed lesion boundary ranges from 0.52 to 0.91 and 0.49 to 0.88, respectively.

III.B. Experiments with clinical images

The largest diameters of the liver lesions ranged from 1.0 cm to 14.4 cm with a mean of 3.3 cm, and the largest diameters of lung nodules ranged from 1.1 cm to 9.8 cm in diameter, with a mean of 3.8 cm. Out of 79 liver lesions, 17 (22%) were on or very close to the boundary of liver and out

of 58 lung nodules, 28 (48%) were adjacent to the pleural wall.

Figure 9 shows the best, worst, and the mean NDCG scores as a function of number of images retrieved for the clinical dataset containing liver lesions [Fig. 9(a)] and lung nodules [Fig. 9(b)] using the reference standard collected from five readers. The NDCG scores indicate good performance of the edge sharpness feature in the task of retrieving images of similar-appearing lesions.

In order to compare the two existing methods that characterize edge sharpness (Gilhuijs and Levman: Sec. II.A.3) to our method that incorporates organ boundary knowledge, we incorporated organ boundary knowledge into the existing methods. We computed the best, worst, and mean NDCG scores using all three methods, and compared the results using Gilhuijs’s and Levman’s features with our margin sharpness feature in the clinical datasets containing liver lesions [Figs. 9(a), 9(c), and 9(e)] and lung nodules [Figs. 9(b), 9(d), and 9(f)]. Our proposed feature outperformed Gilhuijs’s and Levman’s feature in terms of best case, mean, and worst case error over all numbers of images retrieved (K) (Fig. 9). Average performance advantage over Gilhuijs’s method for best case, mean, and worst case NDCG over all K was 1.5%, 3.0% ($p = 0.001$ at $K = 5$, $p = 0.002$ at $K = 10$, paired Wilcoxon test), and 16.2% in the liver data and 2.0%, 5.0% ($p = 0.001$ at $K = 5$, $p = 0.002$ at $K = 10$, paired Wilcoxon test), and 11.7% in the lung data. Average performance advantage over Levman’s method for best case, mean, and worst case NDCG over all K was 1%, 1.5%

TABLE I. Table summarizing the concordance correlation (CC) between the measured gradient and the actual gradient.

	CC mean	CC min	CC max
Proposed method	0.994	0.991	0.996
Gilhuijs’s method	0.900	0.892	0.912
Levman’s method	0.881	0.862	0.894

TABLE II. Table summarizing the concordance correlation between the measured gradient and the actual gradient, after five types of deformations to the lesion margin for each method. Each entry is formatted as: “min CC, [average CC], max CC.”

	Proposed method	Gilhuijs’s method	Levman’s method
Modified by (1)	0.993, [0.995], 0.996	0.801, [0.841], 0.869	0.423, [0.497], 0.552
Modified by (2)	0.988, [0.991], 0.993	0.441, [0.523], 0.586	0.557, [0.679], 0.772
Modified by (3)	0.991, [0.994], 0.996	0.760, [0.804], 0.835	0.603, [0.650], 0.683
Modified by (4)	0.988, [0.991], 0.993	0.908, [0.919], 0.923	0.867, [0.888], 0.902
Modified by (5)	0.991, [0.994], 0.996	0.906, [0.911], 0.927	0.740, [0.775], 0.800

($p = 0.001$ at $K = 5$, $p = 0.001$ at $K = 10$, paired Wilcoxon test), and 15.9% in the liver data and 2.0%, 4.5% ($p = 0.001$ at $K = 5$, $p = 0.003$ at $K = 10$, paired Wilcoxon test), and 12.0% in the lung data. We observe greater improvement for the difficult cases.

Figure 10 shows example results of our system for retrieving images containing liver lesions [Fig. 10(a)] and lung nodules [Fig. 10(b)] with similar margin characteristics. It shows that our system retrieves images with the most similar margin characteristics first, followed by images with less similar margin characteristics.

Fleiss’ kappa was 0.22 (95% CI, 0.21–0.22) for the reference standard for liver lesions and 0.20 (95% CI, 0.19–0.20) for the reference standard for lung nodules, respectively

For the proposed method, there is no statistically significant difference between the NDCG scores obtained using the original lesion margins versus those obtained using the deformed lesion margins in liver and in lung (quantitative results for Schuirmann’s paired TOST are summarized in Table III). However, for Gilhuijs’s and Levman’s method, there is statistically significant difference between the NDCG scores obtained using the original lesion margins versus those obtained using the deformed lesion margins in liver and in lung (quantitative results from paired Wilcoxon tests are summarized in Tables IV and V).

III.C. Parameter sensitivity evaluation

Figure 11 shows how the mean NDCG score for top 10 retrievals for the two datasets varies with different line segment lengths T . From Figs. 11(a) and 11(b) we can see that the mean NDCG score is relatively constant with T , with only slightly worse performance at the extreme edges of scale. We conclude that the segmentation algorithm is insensitive to the parameters predetermined empirically.

III.D. Computation time

For the dataset containing lung nodules, the procedure of margin feature extraction (excluding automatic lung segmentation) required ~ 170 s for all 58 nodules using a PC computer (Intel Core 2 Dual 2.20 GHz with 4 GB of RAM). For the dataset containing liver lesions, given the manual liver segmentation, the procedure required 230 s for all 79 lesions.

This is about 8–9 times slower than the computation time required for Gilhuijs’s and Levman’s feature.

IV. DISCUSSION

In this study, we introduce a method for deriving a quantitative feature to characterize the sharpness of the lesion margin in CT images. This feature may be useful in CBIR applications for retrieving similar-appearing lesions, since lesion margin sharpness is an important aspect reported by radiologists and it varies depending on the type of lesion. Our margin sharpness feature appears promising for assessing the blur and intensity difference across the margin of the lesion; we have shown good performance of this feature in simulated and clinical images. In simulated images, we found good accuracy of our feature in predicting the blur or intensity parameter; our feature predicted the blur and intensity parameters with correlation coefficient greater than 0.95 in all cases (Fig. 7). In contrast, the correlation between the Gilhuijs’s and Levman’s margin features and the blur parameters decays rapidly with the increases in scale parameter, and in that case, the existing features become less discriminative for edges with small amount of blur [Fig. 8(a)]. Because Gilhuijs’ and Levman’s averages several gradient measurements normal to the true edge, these estimates can only be smaller than the gradient at the actual edge, which our method measures by fitting the sigmoid function. This observation explains the decreased bias of our method over the existing methods seen by comparing Figs. 7 and 8. Hence, when there is big intensity difference inside and outside the lesion (i.e., in the clinical dataset containing lung nodules), the proposed feature characterizes the lesion margin better than the existing methods do, as shown in Fig. 9(b).

In clinical images, we have shown that our feature enables retrieving images that are visually similar to a query image (in terms of margin characteristics) with good performance according to NDCG metrics in two different types of lesions, liver and lung (Fig. 9). In Fig. 10(a), when queried with a cyst liver lesion, which is known to have sharply defined edges, the top three retrieved images were all cysts. All retrieved images in the last row (the least similar images) were lesions with other diagnoses. However, we did find substantial spread between the best and worst cases [Fig. 9(a)]. This may be due to a less than ideal reference standard (Fleiss’s Kappa showed only fair inter-reader agreement between readers whose assessments were averaged), to our use of only the margin sharpness as an image feature for CBIR (other image

FIG. 9. Best, worst, and average NDCG plotted for the dataset containing 79 liver lesions, using (a) proposed (c) Gilhuijs's, and (e) Levman's features. Organ border knowledge is used in all methods. In (a), for $K = 5, 10$, and 15 , the combined distance achieved an average NDCG score of 84% , 85% , and 85% , respectively; In (c), for $K = 5, 10$, and 15 , the combined distance achieved an average NDCG score of 78% , 79% , and 81% , respectively; In (e), for $K = 5, 10$, and 15 , the combined distance achieved an average NDCG score of 81% , 82% , and 84% , respectively. Best, worst, and average NDCG plotted for the dataset containing 58 lung nodules, using (b) proposed (d) Gilhuijs's, and (f) Levman's features. In (b), for $K = 5, 10$, and 15 , the combined distance achieved an average NDCG score of 84% , 85% , and 85% , respectively; In (d), for $K = 5, 10$, and 15 , the combined distance achieved an average NDCG score of 75% , 76% , and 77% , respectively; In (f), for $K = 5, 10$, and 15 , the combined distance achieved an average NDCG score of 73% , 75% , and 78% , respectively.

FIG. 10. Examples of retrieval of similar images using one (a) liver lesion and (b) lung nodule as the query image. Query image is in upper left of 8 images in each subfigure, within which rankings go from highest on upper left to lowest on lower right. Due to space limitations, we only show the 3 highest rankings and 4 lowest rankings. Numbers above each image, x (y), show x = computed edge margin similarity score and y = edge margin similarity from the reference standard, both on a scale of 1 to 5 with 5 best.

features such as texture, shape, and other factors would certainly improve CBIR performance) and due to our simplified approach to representing the margin feature using multiple radial lines and averaging the fitted parameters to create a single composite feature. Some lesion types, such as hemangiomas are heterogeneous and, therefore, edge sharpness would not

be expected to be uniform around the entire periphery. Unlike other existing methods, we do take into consideration the situation where a lesion is near the margin of a lesion; we omit the radial lines that would not include the surrounding organ in its entirety, thus eliminating the inclusion of unpredictable

TABLE III. Table summarizing the Schuirmann’s paired TOST between the NDCG scores obtained using the proposed method with original lesion margin and those with deformed lesion margin, at $K = 5$ and $K = 10$, respectively. ($p < 0.05$ to reject the null hypothesis that there is statistically significant difference between the NDCG scores).

Proposed method	K = 5		K = 10	
	Liver	Lung	Liver	Lung
Modified by (1)	0.006	0.003	0.003	0.002
Modified by (2)	0.006	0.004	0.004	0.003
Modified by (3)	0.001	0.005	0.004	0.003
Modified by (4)	0.003	0.002	0.001	0.004
Modified by (5)	0.002	0.001	0.001	0.001

TABLE IV. Table summarizing the paired Wilcoxon test between the NDCG scores obtained using the proposed method with original lesion margin and those with deformed lesion margin, at $K = 5$ and $K = 10$, respectively. ($p < 0.05$ to reject the null hypothesis that there is no statistically significant difference between the NDCG scores).

Gilhuijs’s method	K = 5		K = 10	
	Liver	Lung	Liver	Lung
Modified by (1)	0.020	0.090	0.020	0.010
Modified by (2)	0.005	0.007	0.003	0.001
Modified by (3)	0.006	0.008	0.003	0.004
Modified by (4)	0.008	0.007	0.002	0.005
Modified by (5)	0.030	0.003	0.020	0.001

TABLE V. Table summarizing the paired Wilcoxon test between the NDCG scores obtained using the proposed method with original lesion margin and those with deformed lesion margin, at $K = 5$ and $K = 10$, respectively. ($p < 0.05$ to reject the null hypothesis that there is no statistically significant difference between the NDCG scores).

Levman's method	K = 5		K = 10	
	Liver	Lung	Liver	Lung
Modified by (1)	0.008	0.006	0.003	0.020
Modified by (2)	0.001	0.005	0.009	0.001
Modified by (3)	0.002	0.009	0.001	0.005
Modified by (4)	0.003	0.004	0.003	0.005
Modified by (5)	0.006	0.009	0.001	0.006

structures surrounding the margin of the lesion as a source of error in the margin sharpness feature.

To our knowledge, our work is novel in characterizing the margin of lesions for enabling CBIR. Previous work has been done for characterizing lesion's margin from breast MRI examination to discriminate between malignant and benign lesions.^{22,23} Finally, investigation of our feature in both liver and lung lesions could suggest the utility of this feature in characterizing lesions in other body regions.

V. LIMITATIONS

Our algorithm has several limitations. We assume that there is an accurate segmentation of lesions in order to compute the margin feature. Lesion segmentation was performed by radiologists, and therefore may be subject to interobserver and intraobserver variability. Although automatic lesion segmentation in CT images remains an unsolved image processing task, there is continued substantial progress in certain types of lesions. Various automated segmentation methods for liver lesions and lung nodules have been developed, and some have even recently been deployed in commercial

applications.^{34–39} Since the accuracy of our features is a function of the accuracy of the lesion outline, our method will benefit from the advancement and deployment of the above methods.

A second limitation of our method is that in order for our margin sharpness feature to account for the boundary of the organ in computing the feature, the outline of the organ is required. In some cases, such as in the lung, we were able to automatically obtain this boundary information. However, in other cases, such as in the liver, we required a hand-drawn segmentation. Progress is being made recently in automated organ segmentation, including liver, and thus this limitation will likely lessen over time. Our method could be improved in the future by using automated organ segmentation approaches such as described in Refs. 40 and 41.

A third limitation is that the intensity difference in our margin sharpness feature can be confounded by factors other than the intrinsic lesion characteristics, such as the use of contrast agents or variations in scanning parameters. Standardization in acquisition parameters could reduce this problem, and it is being recognized as being important as quantitative methods such as ours are being developed to evaluate image abnormalities.⁴²

A fourth limitation is the small size of our dataset of 79 liver lesions and 58 lung nodules. However, our initial results in these datasets appear promising and demonstrate the potential value of this feature. Due to the small size of our dataset, some liver lesions were taken from the same patient. The correlation of lesions from the same patient could bias the training of CBIR and reduce its ability to generalize to a large database of cases. However, we also evaluated our method in patients with solitary lung nodules, where this would not present an issue. Moreover, our CBIR results were similar in liver and lung lesion patients (Fig. 9), suggesting that the multiple liver lesions per patient did not bias our results.

A fifth limitation is that it is challenging to arrive at a consensus reference standard from human observers for

FIG. 11. Mean NDCG score for top 10 retrieved images as a function of the length of the normal line segment T for the dataset containing (a) liver lesions and (b) lung nodules.

lesion margin sharpness because of variations amongst viewing radiologists. We used five independent readers and averaged their results to minimize these variations; however, the usual interpretation of Fleiss' kappa indicated only fair agreement amongst the raters.⁴³ We are exploring other methods to obtain more accurate and precise reference standards for evaluating clinical datasets.

A sixth limitation is that clinical data from different CT scanner manufacturers and possibly using different protocols may be highly variable in terms of image quality. The preliminary results presented here, while a solid proof of concept, should be extended to explore dependencies on these factors.

We also note that we used the margin sharpness as the only image feature for CBIR. Our objectives in this paper were to demonstrate a new robust margin sharpness feature and show its utility for retrieving images of lesions with similar margin sharpness. This feature can also be used as a component for a more generalized CBIR that retrieves images that are more generally similar (i.e., other similarities besides margin sharpness, e.g., texture and shape features.¹³)

VI. CONCLUSION

In this paper, we have described a quantitative imaging feature to characterize the sharpness of the margin of lesions on CT images. We have described our system implementation and provided experimental results on clinical and simulated images. In simulated images, we have shown that our feature predicts the gradient parameter more accurately with very high concordance correlation (greater than 0.98 in all cases) than two existing methods, in scenarios with and without the precise lesion boundary. In clinical images, our proposed method outperformed two existing margin characterization methods in terms of average NDCG scores over all K, by 1.5% and 3% in datasets containing liver lesion, and 4.5% and 5% in datasets containing lung nodules. Equivalence testing also showed that our feature is more robust across all margin deformations ($p < 0.05$) than two existing methods for margin sharpness characterization in both simulated and clinical datasets. Thus, our results using this feature are promising and suggest that it is useful for retrieving similar images of lesions, and it will likely be helpful in creating CBIR systems in radiology.

ACKNOWLEDGMENT

This project has been funded from the National Cancer Institute, National Institutes of Health, under Grant Nos. U01CA142555-01 and R01 CA160251.

^{a)} Author to whom correspondence should be addressed. Electronic mail: jiajing@stanford.edu

^{b)} Visiting Professor at the time of research.

¹ S. G. Armato III, M. F. McNitt-Gray, A. P. Reeves, C. R. Meyer, G. McLennan, D. R. Aberle, E. A. Kazerooni, H. MacMahon, E. J. van Beek, D. Yankelevitz, E. A. Hoffman, C. I. Henschke, R. Y. Roberts, M. S. Brown, R. M. Engelmann, R. C. Pais, C. W. Piker, D. Qing, M. Kocherginsky, B. Y. Croft, and L. P. Clarke, "The Lung Image Database Consortium

(LIDC): An evaluation of radiologist variability in the identification of lung nodules on CT scans," *Acad. Radiol.* **14**, 1409–1421 (2007).

² B. J. Hillman, S. J. Hessel, R. G. Swenson, and P. G. Herman, "Improving diagnostic accuracy: A comparison of interactive and Delphi consultations," *Invest. Radiol.* **12**, 112–115 (1977).

³ P. J. Robinson, "Radiology's Achilles' heel: Error and variation in the interpretation of the Röntgen image," *Br. J. Radiol.* **70**, 1085–1098 (1997).

⁴ E. S. Burnside, "Bayesian networks: Computer-assisted diagnosis support in radiology," *Acad. Radiol.* **12**, 422–430 (2005).

⁵ R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.* **40**, 12–17 (2008).

⁶ H. Greenspan and A. T. Pinhas, "Medical image categorization and retrieval for PACS using the GMM-KL framework," *IEEE Trans. Inf. Technol. Biomed.* **11**, 190–202 (2007).

⁷ T. Deselaers, H. Muller, P. Clough, H. Ney, and T. M. Lehmann, "The CLEF 2005 automatic medical image annotation task," *Int. J. Comput. Vision* **74**, 51–58 (2007).

⁸ T. M. Lehmann, M. O. Guld, T. Deselaers, D. Keysers, H. Schubert, K. Spitzer, H. Ney, and B. B. Wein, "Automatic categorization of medical images for content-based retrieval and data mining," *Comput. Med. Imaging Graph.* **29**, 143–155 (2005).

⁹ *Image Clef: Image Retrieval in Clef*, available at <http://imageclef.org/> (April 15, 2008).

¹⁰ W. Hersh, H. Muller, and J. Kalpathy-Cramer, "The ImageCLEFmed medical image retrieval task test collection," *J. Digit. Imaging* (2008).

¹¹ R. Rahmani, S. A. Goldman, H. Zhang, S. R. Cholleti, and J. E. Fritts, "Localized content-based image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 1902–1912 (2008).

¹² M. O. Lam, T. Disney, D. S. Raicu, J. Furst, and D. S. Channin, "BRISC - An open source pulmonary nodule image retrieval framework," *J. Digit. Imaging* **20**(Suppl 1), 63–71 (2007).

¹³ S. A. Napel, C. F. Beaulieu, C. Rodriguez, J. Cui, J. Xu, A. Gupta, D. Korenblum, H. Greenspan, Y. Ma, and D. L. Rubin, "Automated retrieval of CT images of liver lesions on the basis of image similarity: Method and preliminary results," *Radiology* **256**, 243–252 (2010).

¹⁴ M. P. André, M. Galperin, L. K. Olson, K. Richman, S. Payrovi, and P. Phan, "Development of a content-based storage, retrieval and classification system applied to breast ultrasound," *Ultrasonic Image Processing and Analysis Pattern Recognition Letters* (2001).

¹⁵ M. Galperin, "Application of parametric statistical weights in CAD imaging systems," *Proc. SPIE* **5749**, 338 (2005).

¹⁶ M. Galperin, "Statistical analysis to assess automated level of suspicion scoring methods in breast ultrasound," *Proc. SPIE* **5034**, 495–502 (2003).

¹⁷ A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 55–73 (1990).

¹⁸ P. A. Mlsna and N. M. Sirakov, "Intelligent shape feature extraction and indexing for efficient content-based medical image retrieval," in *Proc. of IEEE Southwest Symposium for Image Analysis and Interpretation* (Lake Tahoe, NV, 2004), pp. 172–176.

¹⁹ P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI*, 15 April 2001, Vol. 1, pp. 511–518.

²⁰ J. Z. Wang, G. Wiederhold, O. Firschein, and S. X. Wei, "Content-based image indexing and searching using Daubechies' wavelets," *Int. J. Digit. Lib.* **1**, 311–328 (1997).

²¹ R. Brown, M. Zlatescu, A. Sijben, G. Roldan, J. Easaw, P. Forsyth, I. Parney, R. Sevic, E. Yan, D. Demetrick, D. Schiff, G. Cairncross, and R. Mitchell, "The use of magnetic resonance imaging to noninvasively detect genetic signatures in oligodendroglioma," *Clin. Cancer Res.* **14**, 2357–2362 (2008).

²² K. G. Gilhuijs, M. L. Giger, and U. Bick, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," *Med. Phys.* **25**, 1647–1654 (1998).

²³ J. E. Levman and A. L. Martel, "A margin sharpness measurement for the diagnosis of breast cancer from magnetic resonance imaging examinations," *Acad. Radiol.* **18**, 1577–1581 (2011).

²⁴ W. Dumouchel and F. O'Brien, "Integrating a robust option into a multiple regression computing environment," in *Computing and Graphics in Statistics* (Springer-Verlag, New York, 1991), pp. 41–48.

- ²⁵S. Hu, E. A. Hoffman, and J. M. Reinhardt, "Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images," *IEEE Trans. Med. Imaging* **20**, 490–498 (2001).
- ²⁶J. Wang, F. Li, and Q. Li, "Automated segmentation of lungs with severe interstitial lung disease in CT," *Am. Assoc. Phys. Med.* **36**(10), 4592–4599 (2009).
- ²⁷A. J. Britten, M. Crotty, H. Kiremidjian, A. Grundy, and E. J. Adam, "The addition of computer simulated noise to investigate radiation dose and image quality in images with spatial correlation of statistical noise: An example application to X-ray CT of the brain," *Br. J. Radiol.* **77**, 323–328 (2004).
- ²⁸L. I. K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics* **45**, 255–268, (1989).
- ²⁹J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet* **1**, 307–310 (1986).
- ³⁰J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educ. Psychol. Meas.* **33**, 613–619 (1973).
- ³¹C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval* (Cambridge University Press, Cambridge, 2008).
- ³²S. Siegel, *Nonparametric Statistics for the Behavioral Sciences* (McGraw-Hill, New York, 1956).
- ³³D. J. Schuirmann, "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability," *J. Pharmacokinet Biopharm.* **15**, 657–680 (1987).
- ³⁴J. H. Moltz, L. Bornemann, J. M. Kuhnigk, V. Dicken, E. Peitgen, S. Meier, H. Bolte, M. Fabel, H. C. Bauknecht, M. Hittinger, A. Kiessling, M. Pusken, and H. O. Peitgen, "Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans," *IEEE J. Sel. Top. Signal Process.* **3**, 122–134 (2009).
- ³⁵B. 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).
- ³⁶K. Okada, D. Comaniciu, and A. Krishnan, "Robust anisotropic Gaussian fitting for volumetric characterization of pulmonary nodules in multislice CT," *IEEE Trans. Med. Imaging* **24**, 409–423 (2005).
- ³⁷B. Zhao, L. H. Schwartz, C. S. Moskowitz, M. S. Ginsberg, N. A. Rizvi, and M. G. Kris, "Lung cancer: Computerized quantification of tumor response—Initial results," *Radiology* **241**, 892–898 (2006).
- ³⁸K. Lampen-Sachar, B. Zhao, J. Zheng, C. S. Moskowitz, L. H. Schwartz, M. F. Zakowski, N. A. Rizvi, M. G. Kris, and M. S. Ginsberg, "Correlation between tumor measurement on computed tomography and resected specimen size in lung adenocarcinomas," *Lung Cancer* **75**(3), 332–335 (2012). Epub 2011 Sept. 3.
- ³⁹B. de Hoop, H. Gietema, B. van Ginneken, P. Zanen, G. Groenewegen, and M. Prokop, "A comparison of six software packages for evaluation of solid lung nodules using semi-automated volumetry: What is the minimum increase in size to detect growth in repeated CT examinations," *Eur. Radiol.* **19**, 800–808 (2009).
- ⁴⁰S. Seifert, A. Barbu, S. K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu, "Hierarchical parsing and semantic navigation of full body CT data," in *SPIE Medical Imaging 2009: Image Processing* **7259**, 725902 (2009).
- ⁴¹A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu, "Regression forests for efficient anatomy detection and localization in CT studies," in *Medical Computer Vision, Recognition Techniques and Applications in Medical Imaging* Vol. 6533, edited by B. Menze *et al.* (Springer, Berlin/Heidelberg, 2011), pp. 106–117.
- ⁴²A. J. Buckler, J. L. Mulshine, R. Gottlieb, B. Zhao, P. D. Mozley, and L. Schwartz, "The use of volumetric CT as an imaging biomarker in lung cancer," *Acad. Radiol.* **17**, 100–106 (2010).
- ⁴³J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics* **33**, 159–174 (1977).