**Core samples for radiomics features
that are insensitive to tumor
segmentation: method and pilot study
using CT images of hepatocellular
carcinoma**

Sebastian Echegaray
Olivier Gevaert
Rajesh Shah
Aya Kamaya
John Louie
Nishita Kothary
Sandy Napel

Core samples for radiomics features that are insensitive to tumor segmentation: method and pilot study using CT images of hepatocellular carcinoma

Sebastian Echegaray,^{a,*} Olivier Gevaert,^b Rajesh Shah,^b Aya Kamaya,^b John Louie,^b Nishita Kothary,^b and Sandy Napel^b

^aStanford University, Department of Electrical Engineering, 650 Serra Mall, Stanford, California 94305, United States

^bStanford University, Department of Radiology, 650 Serra Mall, Stanford, California 94305, United States

Abstract. The purpose of this study is to investigate the utility of obtaining “core samples” of regions in CT volume scans for extraction of radiomic features. We asked four readers to outline tumors in three representative slices from each phase of multiphasic liver CT images taken from 29 patients (1128 segmentations) with hepatocellular carcinoma. Core samples were obtained by automatically tracing the maximal circle inscribed in the outlines. Image features describing the intensity, texture, shape, and margin were used to describe the segmented lesion. We calculated the intraclass correlation between the features extracted from the readers’ segmentations and their core samples to characterize robustness to segmentation between readers, and between human-based segmentation and core sampling. We conclude that despite the high interreader variability in manually delineating the tumor (average overlap of 43% across all readers), certain features such as intensity and texture features are robust to segmentation. More importantly, this same subset of features can be obtained from the core samples, providing as much information as detailed segmentation while being simpler and faster to obtain. © 2015 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.2.4.041011]

Keywords: radiomics; CT; image features; segmentation; hepatocellular carcinoma; stability.

Paper 15084SSRR received Apr. 15, 2015; accepted for publication Oct. 14, 2015; published online Nov. 18, 2015.

1 Introduction

The extraction of large quantities of quantitative features from diagnostic medical images is being used for computer-aided diagnosis,^{1,2} screening,^{3,4} and radiomics.^{5,6} Quantitative image features have the potential to provide consistent, nonbiased descriptors of the image being analyzed. Segmenting the region of interest, which might, e.g., represent a tumor, is the first step in being able to extract quantitative features.⁷

Manual segmentation of tumors in CT images is a labor-intensive and time-consuming task.^{8,9} The time taken by a reader to manually segment each slice of a tumor volume limits the amount of cases that can be processed. Also, manually segmented CT images possess intra- and interreader variability,¹⁰ reducing the consistency of any quantitative image descriptor that depends on the shape or position of the segmentation.

Multiple automatic and semiautomatic image segmentation algorithms have been developed to delineate tumors.^{11–15} While these methods increase consistency and reduce the time needed to segment tumors,^{10,16} they are not always able to segment cases that deviate from the norm. Evaluation of these methods is problematic because ground truth is not always available, especially in some difficult cases such as in liver tumors, due to the high variability that human experts have when delineating tumors.

With this in mind, we have decided to test a method that would allow the generation of image feature data from a simple

process to obtain a manually selected subset of the image of a tumor. We first sought to discover features that are robust to changes in the location and shape of the segmentation by calculating the intraclass correlation (ICC)^{17,18} of features obtained from the slice selection and delineations of four readers. Because these features were robust with respect to the segmentation, we explored the possibility of obtaining similar results if we were to ask readers to merely trace a circle, which we call a “core sample,” inside the tumor. We simulated this approach by inscribing the maximum circle inside each reader segmentation and analyzing how the features varied between this core sample and its original segmentation.

In this paper, we present an analysis of the overlap between readers’ segmentations and the robustness of different categories of features with respect to segmentation, along with a core sample technique that could be easily introduced into clinical practice from which a subset of features is consistent with the same features as could be obtained from the detailed segmentation, while being simpler and faster to obtain.

2 Materials and Data

2.1 Data

Following IRB approval by Stanford University’s Research Compliance Office, who waived the requirement for informed consent, we retrospectively selected 29 patients (22 males, 7 females, age range 38 to 85, mean age 64) referred to

*Address all correspondence to: Sebastian Echegaray, E-mail: sechegaray@stanford.edu

Stanford University Medical Center with diagnosed hepatocellular carcinoma (HCC) who had triphasic (arterial, venous, delayed) ($N = 26$) or biphasic (arterial, venous) ($N = 3$) CT scans prior to surgical resection, and whose tumors were subsequently assessed for microvascular invasion. Two patients had two scans within a close interval prior to resection, resulting in a total of 94 scan volumes. Scans were acquired on GE Medical Systems, Siemens, and Toshiba CT scanners with kV, mA, and slice thickness in the range of 80 to 140 kV, 124 to 699 mA, 0.625 to 3 mm, respectively.

2.2 Reader Segmentation

Four radiologists with 6, 10, 12, and 13 years of experience with liver CT independently viewed each image series and selected a centrally located slice along with two additional slices (one more superior and one more inferior, noncontiguous). They manually delineated the tumor in each slice using an open-source annotation tool (ePAD).^{19,20} This created a data set of 282 segmentations per reader, or 1128 segmentations in total. We gave no instructions to the readers about how to segment the slices or which slices to select, except that the first slice chosen had to be centrally located.

2.3 Core Segmentation

From each reader outline, we automatically determined a “core sample” by computing the inscribed circle with maximum circumference (Fig. 1). To obtain the core sample, we first created a grid of equidistant points inside the original segmentation. We then chose the point with the greatest distance from the segmentation boundary. Next, we created a new smaller and finer

Fig. 1 An example of a slice chosen by a reader in the venous phase of one of the cases. The outside trace (blue) is the original segmentation drawn by the reader. The circle inside the segmentation (green) is the automatically determined outline of the core sample.

sampled grid surrounding our candidate point. The process is repeated, until a set precision threshold points is obtained. Finally, to obtain the core sample, we chose the last candidate point and used its position and the minimum distance to the original segmentation boundary as the center and radius of our core sample, respectively. This operation samples the spatial distribution of pixel intensities of the original segmentation while removing information about its shape.

2.4 Image Features

The literature describes many algorithms that can be used to extract quantitative features from regions of interest (ROI) within images.^{21–33} Features can be categorized as measuring intensity, shape, margin, or texture characteristics. Intensity features express statistics of the pixel values within an ROI. Shape features describe the boundary of the ROI. Margin features characterize the transition between the intensity values inside the ROI and the values surrounding it. Lastly, texture features measure the spatial distribution of pixel intensities inside the ROI. A complete list of the features ($n = 745$) used in this study can be found in Appendix A.

2.5 Metrics

2.5.1 Overlap

To analyze the agreement between i ROIs in a given slice, k , we define overlap as the ratio between their intersections and their unions as follows:

$$O_k = \frac{\bigcap_i \text{ROI}_i}{\bigcup_i \text{ROI}_i}. \quad (1)$$

2.5.2 Feature consistency

We used the ICC coefficient to measure the consistency of the features extracted for each segmentation across patients, readers, cores, and slices. The ICC describes how members from the same group resemble each other^{17,18} and has often been used to quantify the consistency of measurements made by different experts.^{34,35} A high ICC shows that a feature is consistent across multiple measurements. This is especially useful when ground truth is not known or hard to find, as is the case here. There are multiple algorithms in the literature to calculate ICC;³⁶ for this study, we used the A-1 method, also known as criterion-referenced reliability expressed as

$$\text{ICC} = \frac{\text{MS}_R - \text{MS}_E}{\text{MS}_R + (k - 1)\text{MS}_E + \frac{k}{n}(\text{MS}_C - \text{MS}_E)}, \quad (2)$$

where MS_R is the mean square for rows (observations), MS_E is the mean square error, and MS_C is the mean square for columns (readers). n and k represent the total number of rows and columns, respectively. In our study, rows represent each slice where features were extracted. Columns represent the different segmentations which can be provided by different readers and/or different methods (e.g., outline, core sample). We used this method as it measures the degree of absolute agreement taking into consideration the systematic variations between readers/methods.

Table 1 Distribution of the absolute distance (in slices) between alternative slices and the initially chosen centrally located slice.

	Mean	SD	Min	25%	50%	75%	Max
Reader 1	6.69	8.07	1	2.00	4.00	8.00	50
Reader 2	11.23	11.33	1	4.00	8.00	14.00	61
Reader 3	15.32	8.07	1	4.00	9.50	19.00	114
Reader 4	7.09	6.51	1	2.00	6.00	9.25	38

3 Results

3.1 Differences in Slice Chosen

Readers varied in how many slices they moved superior and inferior to the first selected slice when choosing the alternative slices. Table 1 shows the absolute distance distribution with respect to the center slice. In this table, we can observe how reader 3 chose slices that were further apart than other readers, while reader 1 usually chose slices that were closer together.

3.2 Overlap

3.2.1 Overlap between readers

The readers' segmentations did not always overlap as they were free to select the slices for segmentation. Table 2 shows the overlap distribution as a function of the number of readers that segmented a particular slice. As expected, as more readers segmented the same slice, the agreement between all of them was diminished. We did not find any particular pair of readers agreeing significantly more [Tukey's Honest Significant Difference Test (Tukey's HSD),³⁷ $p > 0.05$] than the others.

Table 2 Overlap of the manual segmentations as a function of how many readers segmented the same slice. #Slices is the number of slices segmented by at least the number of readers shown in the first column. Overlap is defined in Eq. (1). We show their mean, standard deviation, minimum, and maximum values, and their 25%, 50%, and 75% quartiles.

Readers	#Slices	Mean (%)	SD (%)	Min (%)	25%	50%	75%	Max (%)
2	148	72	16	28	64%	77%	83%	93
3	64	54	16	25	44%	53%	66%	81
4	14	43	14	19	33%	46%	53%	64

Fig. 2 Two samples of 14 slices selected for segmentation by all four readers. Each closed boundary represents a different reader's segmentation. (a) Relatively high overlap (54%) and (b) relatively low overlap (23%).

Table 3 Overlap between each reader's segmentation and its core sample. The overlap between the readers and the core samples is comparable to the overlap between two readers (shown in Table 2).

	Mean (%)	SD (%)	Min (%)	25%	50%	75%	Max (%)
Reader 1	72	12	50	64%	72%	82%	94
Reader 2	79	11	42	73%	82%	88%	95
Reader 3	78	10	54	71%	80%	86%	96
Reader 4	76	10	46	70%	78%	85%	92
All	76	11	42	70%	81%	86%	96

Table 4 Overlap of the segmentations' core samples as a function of how many readers have segmented the same slice. #Slices is the number of slices segmented by at least the number of readers shown in the first column. Overlap is defined in Eq. (1).

Readers	#Slices	Mean (%)	SD (%)	Min (%)	25%	50%	75%	Max (%)
2	148	68	20	12	58%	74%	82%	95
3	64	46	21	5	30%	51%	68%	79
4	14	33	19	5	15%	36%	44%	62

Figure 2 shows examples of slices with relatively high and low overlap.

3.2.2 Overlap between readers and their cores

Table 3 shows the overlap of each reader's segmentation with its core sample. As one can see, the mean overlap is comparable to the overlap between two of the original readers (76% versus 72%). There were no significant differences between readers (Tukey's HSD³⁷ $p > 0.05$).

3.2.3 Overlap between cores

Table 4 shows that the overlap between the readers' cores was lower than the overlap of their manual segmentations (Table 2). While the values are close in the pair comparisons (two readers), the overlap diverges from the overlap between the original segmentations as more readers are added.

3.3 Consistency of Features

3.3.1 Consistency between readers and their cores

We calculated the ICC scores comparing features extracted by each reader using their manual segmentations and their cores. First, we considered only the single centrally located slice chosen by each reader for each phase. Figure 3 plots the distribution of ICC scores for features in four categories (intensity, margin, shape, and texture) for each phase and shows that in all phases, texture and intensity features proved to be the most stable while margin and shape features were the least. Figure 4 ranks the mean ICC scores across phases of the 745 individual features and shows that 584 of them have $ICC > 0.8$. Table 5 lists these ICC scores grouped by subcategory and shows that DB8 wavelets, RMS contrast, Gabor filters, and Haralick features have mean (over all features in each

feature type for all phases) $ICC > 0.8$. We note that existing literature considers ICC values greater than 0.7 to be indicative of strong correlation,^{36,38} so we conclude that these features are highly stable with respect to the segmentation method (outline versus core). We calculated the pairwise correlation for each feature and found that 83% of these correlations were below 0.6, with a median of 0.13, mean of 0.28, and SD of 0.30. These results suggest that the 584 features that show high stability also provide complementary information.

Next, to investigate the effects of a wider variation in slice selection, we computed the ICC for all 81 combinations of readers with selected slices per reader (i.e., four readers, each providing one of their three selected slices) for each phase. Figure 5 plots the distribution of the mean ICC scores per feature subcategory in each phase over the 81 experiments; one can see the same relative stability of feature category independent of the slices selected by the readers. The ICC scores over these 81 experiments of the 745 individual features for each phase are virtually identical to the ones obtained for the centrally located slices, and show that slice selection has little effect on ICC scores that compare features within an outline to features within its core sample.

Please note that in all of the above experiments, only a single slice was used per patient per phase, eliminating any correlation that would be caused by involving multiple slices per patient in the ICC calculation.

3.3.2 Consistency across slices

As in Sec. 3.3.1 above, considering only the data from the single centrally selected slice per reader, we then analyzed the consistency of the features across the different slices chosen by the readers. We sorted the features by their ICC scores and plotted them to observe how many features were consistent in our data set. Figure 6(a) shows that there is high variation among readers

Fig. 3 Comparison of ICC score distributions by feature category in the three phases when only selecting the central slice for each reader per patient per phase: (a) arterial, (b) venous, and (c) delayed.

for each feature when comparing readers using their original outlines. We believe that this is due to the difference in distance between slices that we showed in Table 1, as the least consistent features also had the highest distance between slices (reader 3). It is interesting to note that the spread of the ICC scores between different readers is reduced when we compare the ICC values for the cores [Fig. 6(b)].

4 Discussion

4.1 Core Samples as an Alternative to Border Segmentation

Table 2 shows that expert readers do not always agree about the boundaries of an HCC. This is at least in part due to low-lesion-to-background contrast, heterogeneous density values, and CT noise. Noticing the variability of the boundaries, we decided to explore how much information we could capture with a simple-to-draw core sample. In this study, we simulated drawing core samples by automatically computing them from outlines, but in practice, readers could simply place circles [two-dimensional (2-D)] or spheres [three-dimensional (3-D)] on displayed images much more rapidly than drawing tumor outlines.

Spatial overlap of core samples showed similar statistics when compared to the overlap of original segmentations (Table 3). We also identified that interreader agreement between features extracted from core samples is comparable to those extracted from the original segmentations (Fig. 4) with 584 out of 779 features (75%) showing ICC > 0.8. We note that while Table 5 shows the relative stability among the four feature categories, individual features in each category can be more or less stable than the mean shown. The main advantage of the core samples is their relative ease of acquisition. The main drawback of using core samples instead of the original segmentations is that information about border shape and margin sharpness is lost. This is because all core samples are, by definition, circles and may not conform to the actual margins, while the human-drawn shapes were intended to represent the actual tumor shape and stay close to the margins.

4.2 Stability of Features to Segmentation

A mean spatial overlap as low as 43% when comparing all readers, or 72% when doing pairwise comparisons, shows the importance of having features that are robust with respect to segmentation (Table 2). Therefore, we analyzed which features

Fig. 4 Mean ICC scores across all phases for features calculated between the core samples and their original segmentations, sorted by their scores using the single central slice per patient.

were most stable to the core sampling by calculating the ICC scores between the original segmentations and their cores. We noticed that texture and intensity features remained highly consistent as shown in Fig. 3. This supports the notion that we can capture similar information from an easy to generate core sample, simulated here as the largest circle inside the original outline.

We also looked at the consistency of each feature with respect to the slice chosen. As we can see in Fig. 6(a), feature consistency varies greatly depending on the distance between

the slices chosen. However, this gap is reduced when using core samples as shown in Fig. 6(b). We then looked to see which features were the most stable across different slices and again noticed that intensity and texture features remain highly consistent compared to margin or shape features.

4.3 Limitations

The most obvious limitation of our study is the small cohort of 29 patients. As it is very difficult to obtain outlines of tumors from multiple experienced radiologists, we conducted a small study using four radiologist readers as a first step. However, the data acquired from this pilot study confirmed the large variation in human tracings, and that robust features could be acquired despite this variation. While we make no claims that all of our results will hold up in a larger cohort of patients, these results suggest that such a study is indicated.

One of the limitations of this study is that we only tested for stability but not for clinical relevance of the features. Because we did not yet try to correlate the feature data to any clinical variables, we cannot assert how useful each feature is as a predictor. We also admit that some features that have been previously described as clinically relevant (e.g., shape, margin sharpness) are not captured by core samples. However, our study showed that in HCC, and most likely, in many other tumors, there is high variability in human determination of tumor outlines, leading to increased uncertainty in the features derived from the borders. However, the features that have shown high robustness (Table 5) have been reported to be correlated with clinical variables in previous studies (Gabor filters,^{39,40} Daubechies,⁴¹ and Haralick features^{42,43}). Future studies are needed to further investigate the association of these features with clinical variables.

As observed in Fig. 6, the stability of the features is affected by the distance between the slices being compared (Table 1). This suggests what is generally known, i.e., not all information may be captured by a single slice. Ultimately, segmentation

Table 5 ICC scores between features computed from the core samples and their original segmentation features grouped by method; algorithms shown in Appendix A.

	Number of features	Mean ICC	SD
DB8 wavelets	324	0.97	0.02
RMS contrast	1	0.95	N/A
Gabor filters	40	0.92	0.14
Haralick features	276	0.87	0.12
Intensity statistics	16	0.66	0.29
Michelson contrast	1	0.62	N/A
Haar filters	15	0.29	0.33
Margin scale	36	0.18	0.21
Margin window	36	0.15	0.19
Compactness	1	0	N/A
Roughness	1	0	N/A

Fig. 5 Plot showing category mean ICC scores over all 81 experiments using a single slice per patient per phase calculated comparing core samples to their original segmentations: (a) arterial, (b) venous, and (c) delayed.

Fig. 6 Plot showing the ICC calculated across different slices for each reader, sorted by their scores per reader. (a) ICC scores obtained for the original outlines. (b) ICC scores for the features calculated from the core samples.

should be done in 3-D, and we believe the results shown here will generalize to 3-D core samples. Manually delineating a tumor in 3-D is even more time-consuming and more variable than 2-D outlining. Accurate automated segmentation of these tumors in CT images has, so far, proved elusive. Fortunately, we have shown in this paper that multiple features are stable when selecting a core sample of the tumor, which is less time intensive.

For this study, we simulated the acquisition of core samples by automatic inscription of a circle within the manually outlined tumor. Although we stated that manual acquisition of core samples is much simpler and more rapid to acquire than tumor outlining, we have not tested this explicitly. However, our preliminary results using automatically generated core samples support additional investigation of the stability of features generated within core samples, and the amount of time required to generate them in clinical practice.

Also, we simulated our core samples as the largest circle that could be inscribed in the human tracing. Manually placed core samples might vary in size; however, because we simulated these core samples from four outlines of four independent readers, we feel we have explored any issues that might arise as a function of varied size and placement of core samples.

It is entirely possible that certain tumors may display predictable texture features only near their borders and, in consequence, may be missed by core samples. We could not test for this effect in our small cohort of 29 patients. However, we envision that thoughtful manual placement of core samples by experienced radiologists could compensate for any issues of this type that arises, a topic for further investigation.

Finally, we have only investigated and validated core sampling in the context of CT images of liver tumors. We suspect that our results will generalize to other tumor types and imaging devices, but this claim will have to be specifically evaluated for each application.

Appendix A: Image Features Used in This Study

A.1 Intensity Features

To characterize intensity, we used multiple methods. First, we extracted all the classical statistical descriptors: mean, variance, kurtosis, skewness, entropy, max, and min to characterize the intensity values inside the ROI. Then we measured the contrast, defined as the range between intensity values. We used two methods to measure contrast: the Michelson contrast and the root mean square (RMS) method. The Michelson contrast^{44,45} is defined by

$$\frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}, \quad (3)$$

where $I_{\max}$ and $I_{\min}$ are the maximum and minimum values of the ROI, respectively. This method scales the range by twice the average luminosity, and it is used when dark and light areas are equally probable.

Another method to measure contrast is the RMS contrast^{44,46} defined by

$$\sqrt{\frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (I_{ij} - \bar{I})^2}, \quad (4)$$

where I_{ij} is the intensity value at the ROI location (i, j) , $\bar{I}$ is the average intensity value, and M and N are the height and width of the ROI. The RMS contrast is defined as the standard deviation of the gray intensity of the image after the image has been normalized between $[0, 1]$.

A.2 Margin Features

To characterize the margin of the lesion, we used the method as shown by Xu et al.⁴⁷ In this method, points are selected along the segmentation's boundary, then one calculates vectors, at the selected points, normal to the segmentation boundary. Next, a line segment is fitted at each vector extending a few pixels from each side of the boundary, and image intensity values are sampled along the line segment. Finally, we characterize the sampled values by fitting the sigmoid function

$$f(x; S, W, I_0, x_0) = I_0 + \frac{S}{1 + \exp - \frac{x-x_0}{W}}, \quad (5)$$

where I_0 is the bias intensity, x_0 allows the center of the sigmoid to be different from the center of the line segment, S is the scale of the window where a large S would be indicative of a big intensity change between the inside and outside regions of the segmentation, and W is the width of the window where the width of the window captures the speed of transition between the inner and outer gray values. This regression gives us a parameterization of the boundary at each point in terms of S , W , I_0 , and x_0 .

A.3 Texture Features

To analyze texture, we used multiple algorithms that have been successfully used in the literature: Gabor filters,^{48,49} Haralick features,⁵⁰⁻⁵² and wavelets decomposition.⁵³

Gabor filters are known for representing texture in a similar way to the human vision system,^{54,55} and present multiple advantages for texture extraction and classification.^{56,57} The closed-form equation for the Gabor filter response in 2-D is given by

$$\begin{aligned} \psi(x, y; f_0, \theta) &= \frac{f_0^2}{\pi\gamma\eta} e^{-\frac{f_0^2}{\gamma^2}\tilde{x}^2 + \frac{f_0^2}{\eta^2}\tilde{y}^2} e^{j2\pi f_0 \tilde{x}} \\ \tilde{x} &= x \cos \theta + y \sin \theta \\ \tilde{y} &= -x \sin \theta + y \cos \theta, \end{aligned} \quad (6)$$

where f_0 is the central frequency, θ is the rotation angle, and γ and η are the sharpness along the major and minor axes of the filter. In this work, we extracted the response at five different frequencies and at four different orientations.

Haralick features are a set of statistical measurements taken from a gray-level co-occurrence matrix (GLCM).⁵⁰⁻⁵² These metrics have shown clinical significance multiple times in the medical literature.^{29,30,58-62} A GLCM is defined as the distribution of co-occurring values in an image at a fixed offset. This matrix is defined by

Table 6 List of Haralick features extracted from the gray-level co-occurrence matrix. These features describe the relations of gray values in the ROI within a set offset and orientation. Details of the implementation of each of these features can be found in the references.

Autocorrelation ⁵¹	Contrast ^{50,51}	Cluster prominence ⁵¹
Correlation ^{50,51}	Cluster shade ⁵¹	Dissimilarity ⁵¹
Energy ^{50,51}	Entropy ⁵¹	Homogeneity ⁵¹
Max. probability ⁵¹	Sum of squares ⁵⁰	Sum average ⁵⁰
Sum variance ⁵⁰	Sum entropy ⁵⁰	Difference variance ⁵⁰
Difference entropy ⁵⁰	Info. correlation 1 ⁵⁰	Info. correlation 2 ⁵⁰
Inverse difference ⁵²	Inv. diff. norm. ⁵²	Inv. diff. moment ⁵²

$$C(i, j; \Delta x, \Delta y) = \sum_{p=1}^N \sum_{q=1}^M \begin{cases} 1, & \text{if } I(p, q) = i \text{ and } I(p + \Delta x, q + \Delta y) = j \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where i and j are the row and column indices of the GLCM. Δx and Δy are the fixed offsets in both axes of the image. $I(p, q)$ is the gray level at point (p, q) , and N and M are the height and width of the image. A set of common descriptive statistics is then computed from the GLCM. The complete set of descriptors along with their references is shown in Table 6. To obtain rotation invariant features, we calculate the Haralick features in four directions (0 deg, 45 deg, 90 deg, and 135 deg), aggregate them, and report their maximum, minimum, mean, and standard deviation.

Finally, we also used one-dimensional wavelet feature extraction on the intensity histogram (Haar)^{28,63} and 2-D on the ROI (DB8)²⁷ as shown in Xu et al. pipeline³² to evaluate its robustness.

Acknowledgments

The authors gratefully acknowledge Dr. Daniel L. Rubin for providing access to and support for ePAD, the tool we used to annotate and collect regions of interest for HCC in multiphasic CT scans. This work was supported by the National Institutes of Health grants R01 CA160251 and U01 CA142555.

References

1. K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Comput. Med. Imaging Graph.* **31**(4), 198–211 (2007).
2. H.-P. Chan et al., "Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography," *Med. Phys.* **14**(4), 538–548 (1987).
3. H. Yu et al., "Fast localization of optic disc and fovea in retinal images for eye disease screening," *Proc. SPIE* **7963**, 796317 (2011).
4. J. Zhang and Y. Liu, "Cervical cancer detection using SVM based feature screening," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2004)*, Saint-Malo, France, pp. 873–880, Springer, Berlin, Heidelberg (2004).
5. P. Lambin et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer* **48**(4), 441–446 (2012).
6. V. Kumar et al., "Radiomics: the process and the challenges," *Magn. Reson. Imaging* **30**(9), 1234–1248 (2012).

7. J. Xuan, T. Adali, and Y. Wang, "Segmentation of magnetic resonance brain image: integrating region growing and edge detection," in *Proc. Int. Conf. on Image Processing 1995*, Vol. 3, pp. 544–547, IEEE (1995).
8. W. 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**(1), 63–72 (2006).
9. T. Stammerberger et al., "Interobserver reproducibility of quantitative cartilage measurements: comparison of b-spline snakes and manual segmentation," *Magn. Reson. Imaging* **17**(7), 1033–1042 (1999).
10. L. Hermoye et al., "Liver segmentation in living liver transplant donors: comparison of semiautomatic and manual methods," *Radiology* **234**(1), 171–178 (2005).
11. K. Horsch et al., "Automatic segmentation of breast lesions on ultrasound," *Med. Phys.* **28**(8), 1652–1659 (2001).
12. D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annu. Rev. Biomed. Eng.* **2**(1), 315–337 (2000).
13. D. Pescia, N. Paragios, and S. Chemouny, "Automatic detection of liver tumors," in *5th IEEE Int. Symp. on Biomedical Imaging: From Nano to Macro (ISBI 2008)*, pp. 672–675, IEEE (2008).
14. T. Heimann et al., "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imaging* **28**(8), 1251–1265 (2009).
15. X. Zhang et al., "Interactive liver tumor segmentation from CT scans using support vector classification with watershed," in *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC 2011)*, pp. 6005–6008, IEEE (2011).
16. L. Bergouignan et al., "Can voxel based morphometry, manual segmentation and automated segmentation equally detect hippocampal volume differences in acute depression?" *NeuroImage* **45**(1), 29–37 (2009).
17. G. G. Koch, "Intraclass correlation coefficient," in *Encyclopedia of Statistical Sciences*, pp. 213–217, Wiley-Interscience, Hoboken, New Jersey (1982).
18. A. Caceres et al., "Measuring FMRI reliability with the intra-class correlation coefficient," *NeuroImage* **45**(3), 758–768 (2009).
19. D. L. Rubin et al., "Automated tracking of quantitative assessments of tumor burden in clinical trials," *Transl. Oncol.* **7**(1), 23–35 (2014).
20. "ePAD web-based platform for quantitative imaging in the clinical workflow," <http://epad.stanford.edu> (4 October 2015).
21. F. Gimenez et al., "On the feasibility of predicting radiological observations from computational imaging features of liver lesions in CT scans," in *First IEEE Int. Conf. on Healthcare Informatics, Imaging and Systems Biology (HISB 2011)*, pp. 346–350, IEEE (2011).
22. H. J. Aerts et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat. Commun.* **5**, 4006 (2014).
23. Y. Balagurunathan et al., "Reproducibility and prognosis of quantitative features extracted from CT images," *Transl. Oncol.* **7**(1), 72–87 (2014).
24. S. Basu et al., "Developing a classifier model for lung tumors in CT-scan images," in *IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC 2011)*, pp. 1306–1312, IEEE (2011).
25. X. Gao et al., "Texture-based 3D image retrieval for medical applications," in *IADIS Int. Conf. e-Health*, pp. 101–108 (2010).
26. Y. Rubner and C. Tomasi, "Texture metrics," in *IEEE Int. Conf. on Systems, Man, and Cybernetics 1998*, Vol. 5, pp. 4601–4607, IEEE (1998).
27. I. Daubechies et al., *Ten Lectures on Wavelets*, Vol. **61**, Society for Industrial and Applied Mathematics Philadelphia, Pennsylvania (1992).
28. L. Dettori and L. Semler, "A comparison of wavelet, ridgelet, and curvelet-based texture classification algorithms in computed tomography," *Comput. Biol. Med.* **37**(4), 486–498 (2007).
29. S. Doyle et al., "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *5th IEEE Int. Symp. on Biomedical Imaging (ISBI): From Nano to Macro*, pp. 496–499, IEEE (2008).
30. L. Tesar et al., "Medical image analysis of 3D CT images based on extension of Haralick texture features," *Comput. Med. Imaging Graph.* **32**(6), 513–520 (2008).
31. O. Gevaert et al., "Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data methods and preliminary results," *Radiology* **264**(2), 387–396 (2012).
32. S. A. Napel et al., "Automated retrieval of CT images of liver lesions on the basis of image similarity: method and preliminary results," *Radiology* **256**(1), 243–252 (2010).

33. O. Gevaert et al., "Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features," *Radiology* **273**(1), 168–174 (2014).
34. P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychol. Bull.* **86**, 420–428 (1979).
35. L. Lu and N. Shara, "Reliability analysis: calculate and compare intraclass correlation coefficients (ICC) in SAS," Vol. **14**, Northeast SAS Users Group, Baltimore, Maryland (2007).
36. K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychol. Methods* **1**(1), 30 (1996).
37. J. W. Tukey, "Comparing individual means in the analysis of variance," *Biometrics* **5**, 99–114 (1949).
38. C. P. Dancy and J. Reidy, *Statistics Without Maths for Psychology: Using SPSS for Windows*, Prentice-Hall Inc., Upper Saddle River, New Jersey (2004).
39. F. J. Ayres and R. Rangayvan, "Characterization of architectural distortion in mammograms," *IEEE Eng. Med. Biol. Mag.* **24**(1), 59–67 (2005).
40. A. Ahmadian et al., "A texture classification method for diffused liver diseases using Gabor wavelets," in *27th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMB 2005)*, pp. 1567–1570, IEEE (2006).
41. T. Lambrou, A. Linney, and A. Todd-Pokropek, "Wavelet-based analysis and classification of liver CT," in *Medical Image and Signals IRC Third Annual Plenary Meeting*, pp. 53, Chancellor's Conference Center, Manchester, November 29–30 (2005).
42. C.-M. Wu, Y.-C. Chen, and K.-S. Hsieh, "Texture features for classification of ultrasonic liver images," *IEEE Trans. Med. Imaging* **11**(2), 141–152 (1992).
43. E.-L. Chen et al., "An automatic diagnostic system for CT liver image classification," *IEEE Trans. Biomed. Eng.* **45**(6), 783–794 (1998).
44. H. Kukkonen et al., "Michelson contrast, RMS contrast and energy of various spatial stimuli at threshold," *Vision Res.* **33**(10), 1431–1436 (1993).
45. E. Peli, "Contrast in complex images," *J. Opt. Soc. Am. A* **7**(10), 2032–2040 (1990).
46. P. J. Bex and W. Makous, "Spatial frequency, phase, and the contrast of natural images," *J. Opt. Soc. Am. A* **19**(6), 1096–1106 (2002).
47. J. Xu et al., "Quantifying the margin sharpness of lesions on radiological images for content-based image retrieval," *Med. Phys.* **39**(9), 5405–5418 (2012).
48. S. Marčelja, "Mathematical description of the responses of simple cortical cells," *J. Opt. Soc. Am.* **70**(11), 1297–1300 (1980).
49. J. Ilonen et al., "Image feature localization by multiple hypothesis testing of Gabor features," *IEEE Trans. Image Process.* **17**(3), 311–325 (2008).
50. R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern. SMC-3*(6), 610–621 (1973).
51. L.-K. Soh and C. Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," *IEEE Trans. Geosci. Remote Sens.* **37**(2), 780–795 (1999).
52. D. A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization," *Can. J. Remote Sens.* **28**(1), 45–62 (2002).
53. T. Chang and C.-C. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Trans. Image Process.* **2**(4), 429–441 (1993).
54. J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Am. A* **2**(7), 1160–1169 (1985).
55. M. Haghighat, S. Zonouz, and M. Abdel-Mottaleb, "Identification using encrypted biometrics," in *Computer Analysis of Images and Patterns*, pp. 440–448, Springer Berlin Heidelberg, New York (2013).
56. S. E. Grigorescu, N. Petkov, and P. Kruizinga, "Comparison of texture features based on Gabor filters," *IEEE Trans. Image Process.* **11**(10), 1160–1167 (2002).
57. L.-L. Huang, A. Shimizu, and H. Kobatake, "Robust face detection using Gabor filter features," *Pattern Recognit. Lett.* **26**(11), 1641–1649 (2005).
58. H. Soltanian-Zadeh, S. Pourabdollah-Nezhad, and F. R. Rad, "Shape-based and texture-based feature extraction for classification of microcalcifications in mammograms," in *Proc. SPIE* **4322**, 301–310, International Society for Optics and Photonics (2001).
59. S. Echegaray et al., "Automated analysis of optic nerve images for detection and staging of papilledema," *Invest. Ophthalmol. Vis. Sci.* **52**(10), 7470–7478 (2011).
60. S. Doyle et al., "Automated grading of prostate cancer using architectural and textural image features," in *4th IEEE Int. Symp. on Biomedical Imaging: From Nano to Macro ISBI 2007*, pp. 1284–1287, IEEE (2007).
61. H. Soltanian-Zadeh et al., "Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms," *Pattern Recognit.* **37**(10), 1973–1986 (2004).
62. L. Tesar et al., "3D extension of Haralick texture features for medical image analysis," in *Proc. of the Fourth Conf. on IASTED Int. Conf. (SPPR 2007)*, pp. 350–355 (2007).
63. L. Semler, L. Dettori, and J. Furst, "Wavelet-based texture classification of tissues in computed tomography," in *Proc. 18th IEEE Symp. on Computer-Based Medical Systems 2005*, pp. 265–270, IEEE (2005).

Sebastian Echegaray is a PhD candidate in electrical engineering at Stanford University. He is part of the Integrative Biomedical Imaging Informatics group under his adviser Dr. Sandy Napel. His areas of expertise include multidimensional signal processing and computer-aided diagnosis. He was formerly a staff engineer at VisionQuest Biomedical in New Mexico. He has authored or coauthored six scientific publications and nine conference presentations.

Olivier Gevaert is an assistant professor at Stanford University focusing on developing machine-learning methods for biomedical decision support from multiscale biomedical data. He is an electrical engineer by training with additional training in artificial intelligence, and a PhD in bioinformatics at the University of Leuven, Belgium. He continued his work as a postdoc in radiology at Stanford. He now is leading a lab on multiscale data fusion at the Department of Medicine at Stanford.

Rajesh Shah is a clinical assistant professor of radiology at Stanford University and chief of interventional radiology at the VA Palo Alto Health Care System. He is an interventional radiologist with a particular interest in diagnosis and treatment of cancer. He performs a wide variety of procedures including chemoembolization, radioembolization, and ablation of tumors. His research interests include the application of big data to medicine and clinical trials.

Aya Kamaya, MD, is an associate professor of radiology at Stanford University and co-director of the Body Imaging Fellowship. Her area of expertise is in abdominal imaging, oncologic imaging, and ultrasound. She has authored over 45 scientific papers, 80 conference presentations, and is the lead author of the textbook *Diagnostic Ultrasound: Abdomen and Pelvis* (Elsevier), as well as lead author for RadPrimer Ultrasound Case Challenges. She is a fellow of the SAR and SRU.

John Louie has been an interventional radiologist at Stanford University for the last 9 years. He specializes in interventional oncology, portosystemic shunts, and cone beam CT. Interventional oncology includes liver-directed therapies such as chemoembolization and radioembolization. He is also a passionate educator, training over 40 fellows and over 100 residents.

Nishita Kothary is an associate professor in the Department of Radiology at Stanford University. Her primary interest is in hepatocellular carcinoma and development of emerging platforms to diagnose and treat this disease. As Stanford faculty, her work includes the use cone-beam CT for image-guided oncologic interventions. Her interest in emerging technologies and her clinical expertise makes her ideally suited as a physician scientist, with collaboration with basic scientists and clinicians in the field.

Sandy Napel received his BSES from SUNY Stony Brook in 1974, and his MS and PhD in electrical engineering in 1976 and 1981, respectively, from Stanford University. He was formerly VP of engineering at Imatron Inc, and is currently professor of radiology and, by courtesy, of electrical engineering and medicine (biomedical informatics research) at Stanford University. He co-leads the Stanford Radiology 3-D and Quantitative Imaging Lab and Radiology Department's Section on Integrative Biomedical Imaging Informatics, where he is developing techniques for linkage of image features to molecular properties of disease.