Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Content-based image retrieval in radiology: analysis of variability in human perception of similarity

Jessica Faruque
Christopher F. Beaulieu
Jarrett Rosenberg
Daniel L. Rubin
Dorcas Yao
Sandy Napel

SPIE.

Content-based image retrieval in radiology: analysis of variability in human perception of similarity

Jessica Faruque,[a,*] Christopher F. Beaulieu,[b] Jarrett Rosenberg,[c] Daniel L. Rubin,[d] Dorcas Yao,[e] and Sandy Napel[f]
[a]Stanford University, Department of Electrical Engineering, 350 Serra Mall, Stanford, California 94305, United States
[b]Stanford University Medical Center, Department of Radiology, 300 Pasteur Drive, Room S078, MC 5105, Stanford, California 94305, United States
[c]Stanford University, Department of Radiology, Lucas MRS Imaging Center, 1201 Welch Road, Room P-280, Stanford, California 94305-5488, United States
[d]Stanford University, Departments of Radiology and Medicine (Biomedical Informatics), Richard M. Lucas Center P285, 1201 Welch Road, Stanford, California 94305-5488, United States
[e]Stanford University, Department of Radiology, 3801 Miranda Avenue, Palo Alto, California 94304-1290, United States
[f]Stanford University, Department of Radiology, James H. Clark Center, 318 Campus Drive, W3.1, Stanford, California 94305-5441, United States

Abstract. We aim to develop a better understanding of perception of similarity in focal computed tomography (CT) liver images to determine the feasibility of techniques for developing reference sets for training and validating content-based image retrieval systems. In an observer study, four radiologists and six nonradiologists assessed overall similarity and similarity in 5 image features in 136 pairs of focal CT liver lesions. We computed intra- and inter-reader agreements in these similarity ratings and viewed the distributions of the ratings. The readers' ratings of overall similarity and similarity in each feature primarily appeared to be bimodally distributed. Median Kappa scores for intra-reader agreement ranged from 0.57 to 0.86 in the five features and from 0.72 to 0.82 for overall similarity. Median Kappa scores for inter-reader agreement ranged from 0.24 to 0.58 in the five features and were 0.39 for overall similarity. There was no significant difference in agreement for radiologists and nonradiologists. Our results show that developing perceptual similarity reference standards is a complex task. Moderate to high inter-reader variability precludes ease of dividing up the workload of rating perceptual similarity among many readers, while low intra-reader variability may make it possible to acquire large volumes of data by asking readers to view image pairs over many sessions. [DOI: 10.1117/1.JMI.2.2.025501]

Keywords: liver; perception; content-based image retrieval; reference standards; computed tomography; similarity.

Paper 15011R received Jan. 27, 2015; accepted for publication Mar. 10, 2015; published online Apr. 3, 2015.

1 Introduction

Liver cancer is a leading cause of cancer mortality, and early detection of cancer using radiological images plays a crucial role in improving survival rates.[1] However, early detection is challenging due to inter-reader variability in image interpretation and the need for more efficiency in interpreting large volumes of imaging data resulting from increased screening rates.[2–4]

Predictive and learning techniques for assisting in radiological decision making are becoming increasingly sophisticated and may improve the accuracy and efficiency of cancer diagnosis.[5,6] One technique, content-based image retrieval (CBIR), involves the presentation of images containing visually similar lesions alongside images requiring diagnosis in order to improve diagnostic accuracy and efficiency.[7–9] Studies involving CBIR systems for clinical decision support demonstrate that they may be useful in assisting diagnosis.[10–12]

A recurring challenge in building a robust CBIR system is obtaining an accurate reference set (or "gold standard") for training and validating the system. Even though major strides have made in image feature analysis and prediction algorithms, obtaining training and testing data that accurately reflect the higher-level perceptual judgments necessary for diagnosis remains a major obstacle to progress.[13–21]

The study of medical image perception is an ongoing task.[22–25] Recent publications about medical image similarity have focused on mammograms. Many of these studies showed that asking several radiologists and nonradiologists to rate similarity in pairs of mammograms may be a feasible way to acquire similarity data, particularly when readers' ratings are averaged.[26,27] Recent work also involves using artificial neural networks to impute similarity ratings, showing that the neural networks produced ratings with variability that may be comparable with variability between radiologist readers.[28,29] Similar work was also performed with lung nodules using a classifier that used perceptual similarity measures to determine if the lesion was benign or malignant.[30]

Some of these techniques focus on predicting similarity in image features rather than overall similarity between images, which is a much more difficult task.[31] Also, while many of these characterize interoperator variability in evaluating similar images, these techniques lack scalability, making it difficult to obtain sufficient data for large databases. When investigating image similarity, asking readers to view every pairwise combination of images in a large database is a daunting and time-consuming

*Address all correspondence to: Jessica Faruque, E-mail: jessica.faruque@alumni.stanford.edu

2329-4302/2015/$25.00 © 2015 SPIE

task. Additionally, many of these studies involve mammography with a benign or malignant classification or other binary classifiers, not addressing cases with multiple diagnoses or appearances. Finally, some of these studies do not address intra-reader agreement and the repeatability of a reader's subjective ratings of image similarity. Thus, considerable work still remains to be done in the topic of medical image similarity.[32]

In previously published work, we developed a scheme for finding a reference standard using three readers' ratings of a variety of image features from images viewed individually and imputing pairwise similarity ratings from these values.[33] While this technique has a great deal of promise as a method for developing a reference standard that may be used for large databases, we found that moderate to high inter-reader variability makes it difficult to combine readers' ratings. We then developed a model that predicts inter-reader agreement for studies with more readers. This model predicts that inclusion of 10 or more readers may overcome some of the issues related to inter-reader variability.[34]

In this paper, we investigate data regarding the similarity of features of images of liver lesions seen at computed tomography (CT) from 10 readers in order to quantify various characteristics of these data such as the distributions of the ratings, intra- and inter-reader agreements, and relevance of specific image features. Additionally, we determine if the variability in these data makes different paradigms of perceptual similarity data collection feasible. By quantifying many characteristics of the perception of image similarity in liver lesions, this work will assist with a wide range of applications in CBIR and medical decision support.

2 Methods

Institutional approval was obtained to allow use of the images, collection of observer data, and the analyses performed in this project. The readers in the study provided informed consent prior to participation.

2.1 *Data Collection*

Image selection and image pair generation: A radiologist selected 53 DICOM images of liver lesions from 53 different patients (27 male, 26 female, age range 24 to 90, acquired on May 2001 through December 2009) containing focal liver lesions comprising a variety of diagnoses (Table 1) from our PACS system, which were deidentified in a HIPAA-compliant manner prior to use. These images were CT scans in the portal venous enhancement phase reconstructed on a 512 by 512 grid with slice thickness ranging from 2 to 10 mm, with the majority being 5 mm. The radiologist outlined each lesion with a rectangular region of interest.

Given 53 images, there are 1378 unique pairings (computed combinatorially as "53 choose 2"). We selected a subset of all pairwise combinations of the 53 image pairs for evaluation. This was done for multiple reasons, with the first being that it is quite time consuming for readers to evaluate 1378 pairs of images. Second, since the goal of this project is for CBIR for similar images, we selected image pairs that are somewhat similar to each other, which are more relevant for achieving our goal.

Based on this reasoning, we generated a set of 136 image pairs as follows. First, two radiologists (one of whom also selected the images) who were not participants in the study, by consensus, divided the 53 images into nine groups (Table 2), with each group containing four to nine images. Each of the

Table 1 Diagnoses of the 53 CT images of focal liver lesions used in the study.

Diagnosis	Number of images
Metastasis	15
Hepatocellular carcinoma	10
Hemangioma	5
Abcess	4
Carcinoid	2
Cholangiocarcinoma	2
Cyst	2
Focal nodular hyperplasia	2
Gastrointestinal stromal tumor	2
Adenoma	1
Confluent hepatic fibrosis	1
Glomus tumor	1
Inflammatory pseudotumor	1
Lymphoma	1
Sarcoma	1
Two or more diagnoses	3
Total	53

Table 2 The 53 images divided into 9 groups of images.

Image group no.	No. of images	Description of visual appearance
1	6	Hypodense, well-defined, and homogeneous
2	4	Isodense, homogeneous, and somewhat well-defined
3	5	Slightly hypodense, heterogeneous, and moderately well-defined
4	5	Heterogeneous, hypodense, and moderately well-defined
5	7	Hypodense, homogeneous, and moderately well-defined
6	6	Ill-defined, heterogeneous, and hypodense
7	5	Heterogeneous and well-defined
8	7	Mixed densities with nodular areas of hyperdensity
9	8	Well-defined, mixed densities, loculations (compartments), bands
Total	53	

groups was defined by phrases describing the visual characteristics of the lesions within the images. For example, the first group was described as "hypodense, well-defined, and homogeneous." Each group contained images that were similar to each other in visual appearance, though not all the images in a group had the same diagnoses. Next, we generated a set of all pairwise combinations of the images within each of the nine groups. For example, in Group 1, we obtained 15 image pairs by generating all pairwise combinations of all 6 images in the group. Finally, we combined all the pairs generated from each of these groups into a set of 136 image pairs. This resulted in the 136 selected image pairs being more similar to each other than the pairs not included. However, the 136 pairs still varied substantially in the amount of similarity between the images. The participants were not told about the image selection process or the fact that we chose a specific subset of all possible image pairs.

Study participants: Ten readers (four radiologists not involved in image selection and presentation and six nonradiologists) participated in this study. We included both radiologists and nonradiologists in order to determine if the two groups differed in their perception of similarity. Of the radiologists, one was a fifth year fellow and three were the faculty with 10, 18, and 31 years of experience, respectively. All the readers were compensated with a $5 gift card for their participation.

Image features and reader training: We asked readers to provide a numerical rating for overall visual similarity between the lesions presented in each pair of images on a continuous scale of 1 (for least similar) to 9 (for most similar). We also asked readers to evaluate lesion similarity in five separate image features (Table 3) that two radiologists selected by consensus. There were two reasons we asked readers to perform this task: (1) to determine how the perception of overall similarity and the perception of feature similarity are related, and (2) to investigate if inter- and intra-reader variability are lower in ratings of specific imaging features. Figure 1 shows the training examples that each reader viewed prior to rating similarity in each feature.

Image pair randomization and repetition: We used the following paradigm for image pair presentation to allow determination of both inter- and intra-reader agreements as well as the effects, if any, of the ordering of the image pairs. First, we withheld 15 out of the 136 pairs of images. We randomized the order of the remaining 121 pairs of images and divided them into three groups of approximate thirds, containing 40, 41, and 40 pairs, respectively. The withheld 15 pairs were added into each of these three groups so that the groups contained 65, 66, and 65 pairs, respectively, so that intra-reader variability could be estimated. Next, the ordering of the pairs within each of these groups was randomized. Finally, we assembled each of these thirds sequentially to create a set of 166 image pairs (including repetitions) that were presented to readers. This process resulted in three presentations of the 15 pairs of images that were viewed

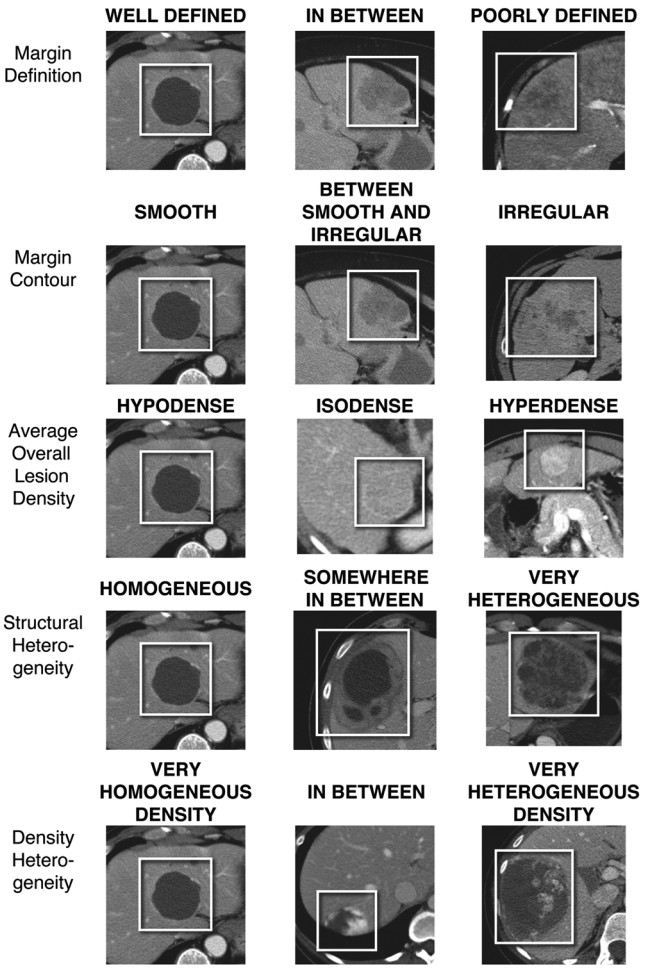

Fig. 1 Training examples presented to the readers prior to performing the study.

Table 3 The five features evaluated by the study participants.

Feature	Description
Margin definition	How sharp or blurry the margin of the lesion was against the surrounding normal liver tissue.
Margin contour	The smoothness or irregularity of the lesion's outer contour or the shape of the lesion boundary against the surrounding normal liver tissue. Readers were asked to disregard the lesion-to-liver contrast and the sharpness of the margin.
Average overall lesion density	The attenuation (dark or bright) relative to the surrounding normal liver tissue, ranging from hypodense to hyperdense. Readers were asked to disregard densities of any rim or capsules present in the lesion. If the lesion was heterogeneous, they were instructed to estimate what the mean density would be if they combined component densities.
Structural heterogeneity	The complexity of the lesion structure such as the "number of compartments" within a lesion.
Density heterogeneity	The number of different densities that is visible in the lesion. This could include density of enhancement.

multiple times by the readers. All readers viewed the same 166 image pairs, and the left-right ordering of the two images in each pair was randomized to reduce the possibility of bias from this ordering.

Image pair presentation: The images were presented on a web-based graphical user interface generated using the Qualtrics survey software (Qualtrics, Provo, Utah). Before viewing the images, the readers were first asked to answer some questions such as whether or not they were radiologists, and if a radiologist, their training levels. Next, they viewed a brief statement explaining the task and the rating scheme.

For consistency in image presentation, all readers viewed the images at the same workstation, which consisted of an Apple MC914LL/A monitor with a 27-in. display and a screen resolution of 2560 by 1440 pixels. All the images were set to have a standard CT liver window (40/400 HU window/level). To prevent the feature similarity ratings from biasing the readers' ratings of overall similarity, readers first rated only overall similarity between the lesions in the image pairs (Fig. 2). Following rating overall similarity, readers rated similarity in each feature.

To avoid reader fatigue and reduce bias, the study was divided into two sessions. While six separate sessions would have been ideal, this proved difficult to schedule and would have reduced participation in the study. During the first session, readers rated overall similarity and similarity in two features. During the second session, readers rated similarity in the remaining three features. The presentation order of the features for each reader was randomized, so each of the readers rated different subsets and different orderings of the features in each session.

Readers' comments: Immediately after rating overall similarity, we asked readers a set of subjective questions about the basis by which they determined similarity, what lesion features they used in making similarity judgments, and, for the radiologist readers, what imaging features they primarily use when making diagnoses involving liver lesions. After readers completed all the tasks, we asked them about the difficulty level of the study and what challenges they faced, if any.

2.2 *Data Analysis*

We performed the following analyses to characterize various properties of these data such as intra- and inter-reader variability.

Distributions of the ratings: First, we viewed histograms of the readers' ratings to see what the distributions of ratings looked like and if readers' ratings followed similar distributions.

Intra-reader agreement: Next, we computed intra-reader agreement using a quadratically weighted Kappa statistic between the first and second, first and third, and second and third instances of the 15 repeated pairs for each reader.[35] We used a Wilcoxon signed-rank test to determine if any differences existed between the three different sets of Kappas. For these values of Kappa and all following, we used the scaling detailed by Landis and Koch[36,37] to determine the agreement level. We also used a Wilcoxon signed-rank test to determine if any statistically significant difference between the Kappa values for radiologists and nonradiologists existed.

Inter-reader agreement: To determine reader agreement, we computed a quadratically weighted Kappa statistic between every pair of readers. Again, we used a Wilcoxon signed-rank test to determine if any statistically significant difference between the Kappa values for radiologists and nonradiologists existed. As before, we performed these computations and tests for both the feature and overall similarity ratings.

We viewed image pairs in which inter-reader agreement was either very high or very low in overall similarity to see if there were any distinguishing characteristics. We computed the Euclidean distance between the 10 readers' ratings of overall similarity for each image pair to measure the inter-reader agreement for that pair.

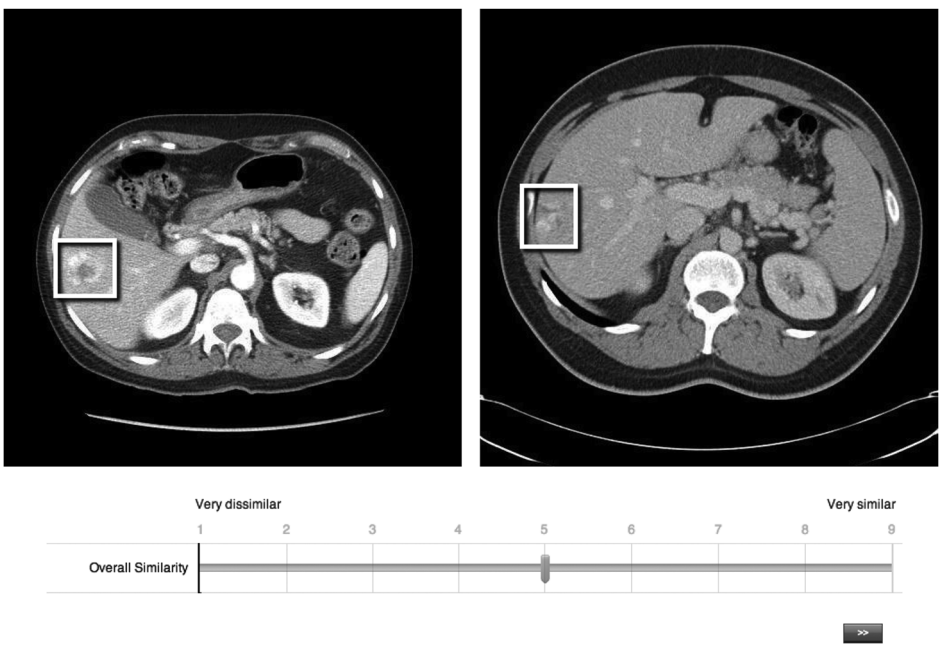

Fig. 2 Screen capture of the graphical user interface (GUI) presented to readers for overall similarity. The GUI for each of the features was similar, with the words "overall similarity" replaced by the name of the feature.

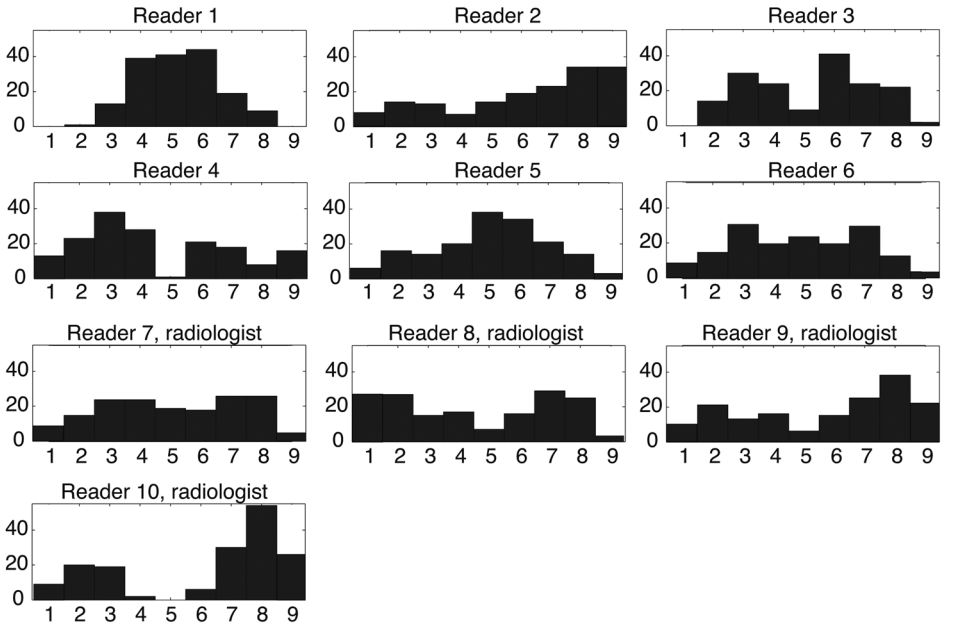

Fig. 3 Distributions of ratings of radiologist and nonradiologist ratings for overall similarity between the lesions in the images. The *y*-axis shows the number of occurrences at each of the values on the 9-point rating scale.

3 Results

Distributions of the ratings: Distributions of ratings of all four radiologists' and four of the six nonradiologists' ratings of overall similarity appeared to be bimodal distributions (Fig. 3). For the remaining readers, the distributions appeared to be Gaussian. The feature similarity ratings also showed mostly bimodal distributions that were similar to the distributions of ratings for overall similarity.

Intra-reader agreement: We report median values of Kappa since the distribution of Kappa values is often skewed. The median Kappa values for intra-reader agreement for overall similarity between the first and the second, first and third, and second and third presentations of the 15 repeated pairs were 0.78, 0.72, and 0.82, respectively (Fig. 4). For all the feature similarities and image repetitions, the median intra-reader agreement ranged between 0.57 (for agreement between the second and third presentations for the density heterogeneity feature) and 0.86 (for agreement between the second and third presentations for the average density feature). Hypothesis testing for differences between the radiologists and nonradiologists did not achieve statistical significance (p-values ranged from 0.08 to 0.82).

Inter-reader agreement: The median Kappa value for inter-reader agreement values for overall similarity was 0.39 (Fig. 5).

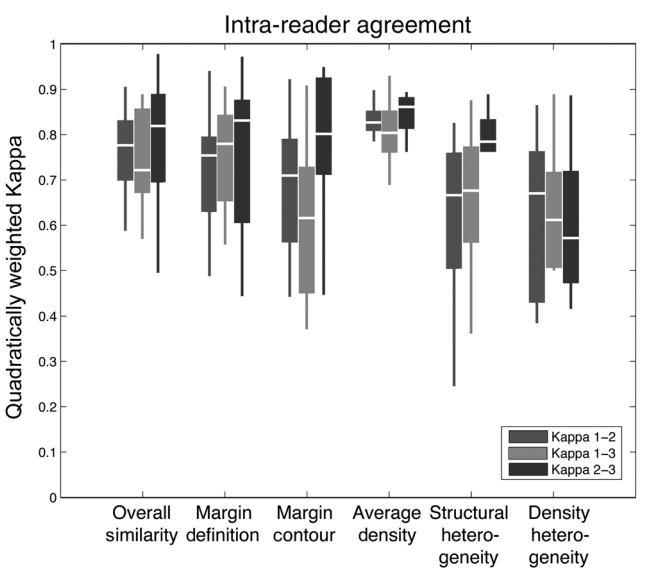

Fig. 4 Intra-reader agreement (Kappa) between the first and second, first and third, and second and third presentations of image pairs for all the readers. Plots show the median (white line), minimum and maximum (ends of the whiskers), and the first and third quartiles (ends of the boxes).

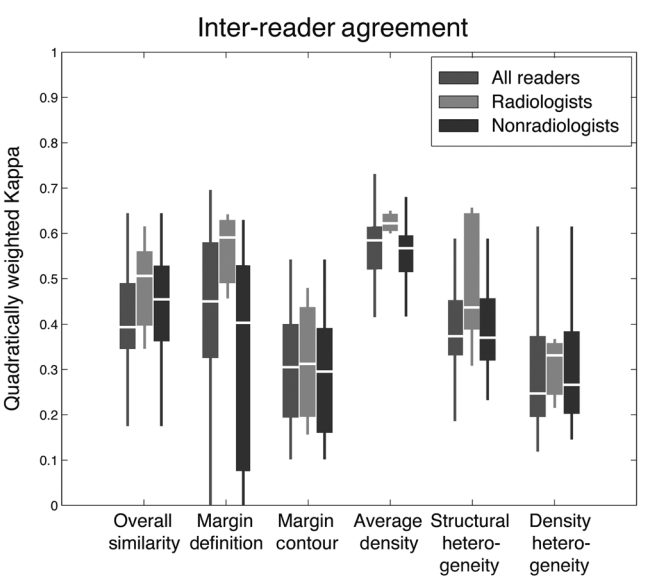

Fig. 5 Inter-reader agreement for overall similarity and similarity in each of the five features. Plots show the median (white line), minimum and maximum (ends of the whiskers), and the first and third quartiles (ends of the boxes).

For only radiologists and only nonradiologists, respectively, the median Kappa values were 0.51 and 0.45. For the image feature similarity ratings, the median inter-reader agreement for all the readers ranged from 0.24 (for the density heterogeneity feature) to 0.58 (for the average density feature). For the radiologists' ratings of image features, the median values ranged from 0.31 (for the margin contour feature) to 0.62 (for the average density feature); for the nonradiologists, the median values ranged from 0.26 (for the density heterogeneity feature) to 0.57 (for the average density feature). Hypothesis testing for differences between the inter-reader agreement in radiologists, nonradiologists, and all readers did not achieve statistical significance for overall similarity and each of the features (p-values ranged from 0.08 to 0.42) using a two-sided Wilcoxon test.

We viewed the image pairs in which the readers' ratings for overall perceptual similarity agreed the least and the most as measured by the Euclidean distances between the ratings (Fig. 6). The five image pairs with the greatest agreement, or the least Euclidean distance between the ratings, all appeared to be pairs in which readers thought that the two lesions were very similar or very dissimilar. In the five image pairs with the least agreement, or the greatest Euclidean distance between the ratings, readers' ratings for each pair spanned the entire scale. Of these five pairs, one pair was the second pair presented in sequence to odd numbered readers (and the 165th pair presented

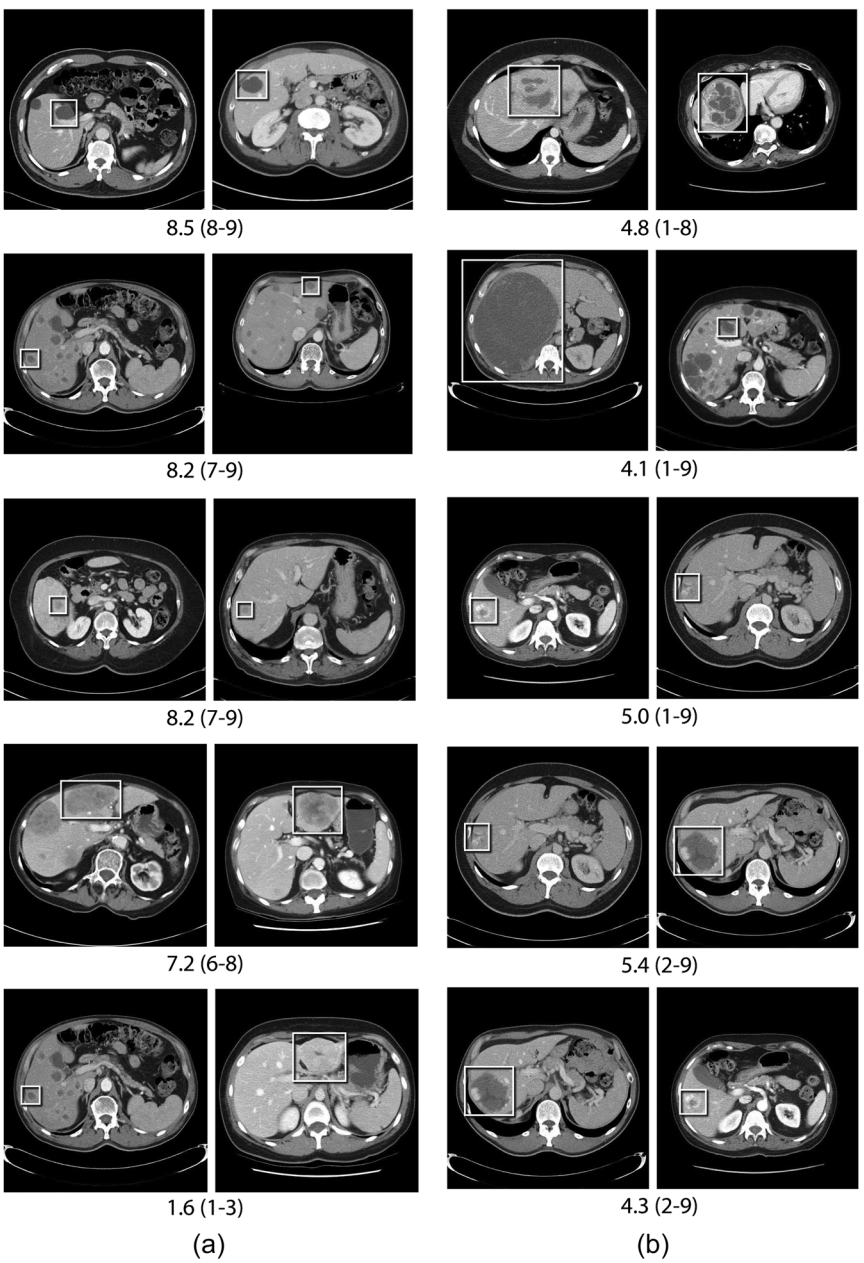

Fig. 6 (a) Five pairs of images in which agreement between the readers was the greatest, and (b) five pairs of images in which agreement between the readers was the least. High values indicate that readers thought lesions were very similar, whereas low values indicate that readers thought the lesions were very different. Mean similarity ratings are shown below each image, with the ranges in parentheses.

to even numbered readers), and another pair was the 166th pair presented to the odd numbered readers (and the very first pair presented to even numbered readers).

Readers' comments: Following completion of rating overall similarity in the images, multiple readers responded that they used the following lesion features for their similarity judgments:

- Lesion texture
- Lesion edge sharpness
- Density (or brightness or intensity)
- Contrast
- Lesion size.

4 Discussion

4.1 *Data Collection*

Our technique of selecting a subset of pairs of already somewhat similar images allowed readers to complete the study in a matter of hours. One of the challenges with this technique, however, is that the results may change when highly dissimilar images are also included. Studies that include images which are more dissimilar to each other, such as a random sampling of all possible pairs, may be used to investigate this. Additionally, the process of grouping the 53 images for this study into groups that were somewhat similar to each other is in itself a task that may be subjected to high levels of variability. Thus, this work could also benefit from further studies that evaluate the repeatability and variability of performing this task.

4.2 *Data Analysis*

Distributions of ratings: One possible reason for the mostly bimodal distribution of the ratings is that the readers may perceive similarity between images as either somewhat similar or somewhat dissimilar and then refine further within each category. Since the image pairs presented were within groups of images that were chosen to be somewhat similar to each other, we hypothesized that a larger fraction of the ratings would be on the end of the scale corresponding to images that were considered very similar. However, most readers used the entire 9-point rating scale and did not seem to provide more ratings near the upper end of the scale corresponding to high similarity. It is interesting that in a previous study that presented images that were assumed to be very dissimilar as well as images that were assumed to be similar to each other, readers' ratings of similarity were also bimodally distributed across the entire scale.[33] It is thus possible that the selection of any subset of image pairs may nonetheless result in bimodally distributed ratings of similarity as the readers calibrate their ratings based on the image pairs that they see.

Another notable observation is that the distributions of all the radiologist readers' ratings were bimodal, whereas two nonradiologists' distributions of ratings were not. One possible explanation for the radiologists' better consistency is that they have years of training in medical image interpretation. Another explanation is that the distributions of the ratings may cluster into different groups, with bimodal distributions being the largest group.

Intra-reader agreement: The median values of intra-reader agreement for the overall similarity ratings were above 0.8, which is considered excellent according to the scale by Landis and Koch.[35,36] For the feature ratings, the intra-reader agreement for average density appeared to be the highest, with excellent agreement, followed by margin definition, also with excellent agreement. The median values for intra-reader agreement for margin contour, structural heterogeneity, and density heterogeneity were above 0.6, and all values except the median agreement between the second and third presentations of the density heterogeneity feature were above 0.6, which is considered substantial agreement. Since statistically significant differences were not noted between intra-reader agreement in the image pairs, there may not be a "learning curve" for readers to decide how to rate similarity in images as hypothesized. However, the small sample size (15 repeated pairs) may not have been sufficient to reveal this effect.

One technique for obtaining similarity ratings for large image databases involves dividing the tasks into multiple sessions over longer periods of time. Since the levels of intra-reader agreement we measured are generally high, this technique may thus be feasible.

Inter-reader agreement: There was fair to moderate inter-reader agreement for overall similarity between readers. For the features, inter-reader agreement varied from fair, with the lowest and highest agreement levels being density heterogeneity and average density, respectively.[35] The median values for inter-reader agreement for each of the features were generally lower than the values for intra-reader agreement, which is reasonable since readers are expected to agree more with themselves than others. There was no statistically significant difference in inter-reader agreement between radiologists and nonradiologists, which may be because radiologists are trained to make diagnoses rather than to evaluate similarity between images. Also, a significant difference may not be apparent because of the small number of radiologist readers.

Inter-reader agreement may be low for some features because different readers may have different ways of interpreting similarity in those features. For example, with structural heterogeneity and density heterogeneity, readers may disagree on a pair of lesions because some readers may interpret a lesion as homogeneous with a fine texture, whereas other readers may interpret the same lesion as having many different pieces and thus very heterogeneous.

Our study showed rectangular ROIs around each tumor to distinguish it from any other tumors that might have been on the image and asked radiologists to rate features of what they perceived to be the tumor. Within the rectangle, perceptions of whether a region is part of the tumor may vary among radiologists and may contribute to inter-reader variability.

Our considerations when selecting image pairs included (a) selecting a number of pairs that could be viewed by the participants in a reasonable amount of time, and (b) selecting a subset of pairs that is somewhat similar to each other. To achieve this, we divided the images into groups and selected image pairs for comparison from within each group. However, this technique has its limitations, such as inadequate representation of all possible pairs, and not obtaining information about highly dissimilar pairs.

In the image pairs with the least and most Euclidean distance between the ratings, the pairs with the least distance (or most agreement) were the pairs that readers all rated as either very similar or very dissimilar. This is expected, since readers may agree more with each other when rating similarity at the extreme ends of the spectrum. The image pairs with the most

distance (or least agreement) appear to contain images that are more ambiguous in interpretation. For example, in images in which a region of the lesion was isodense with the surrounding liver, some readers may have considered this region to be part of the lesion and others may have not, resulting in discrepancies in the similarity ratings. Also, two of the five image pairs with the greatest Euclidean distance were among the very first pairs that some readers viewed during the study. The ratings for these pairs may have high variability because readers had not yet viewed enough pairs to calibrate their ratings relative to the other presented pairs. Approaches for reducing both intra- and inter-reader variability may include presenting more training examples prior to beginning the similarity evaluation task. Additionally, there are a number of other distance metrics that could be used in the future instead of the Euclidean distance metric, each with their advantages and disadvantages.

One technique of obtaining similarity ratings for large datasets involves assigning a different subset of images to each reader and patching the results together into a similarity matrix. Since the results here show that the inter-reader agreement was only moderate in some cases, this technique may be challenging to implement without first investigating techniques for improving inter-reader agreement.

Responses to survey questions: In the survey responses, many of the imaging features that the readers noted were useful in making both diagnoses and image similarity judgments were also the features that we selected for the study tasks, which provided good verification that these were indeed relevant features for analysis. These included features such as texture, sharpness, and density. However, several readers also responded that the lesion size, which we asked readers to disregard, was also a factor in their decision making. Based on this, it may be useful in future studies to ask readers to evaluate lesion size in addition to existing features. While most of the readers felt that the image texture was important, high inter-reader variability existed in the ratings of structural heterogeneity and density heterogeneity.

In conclusion, our results show that the medical image similarity perception is a complex visual task that requires rigorous attention to training, experimental detail, and careful attention to intra- and inter-reader effects. Our results also show that determining whether CT liver image similarity can be quantified well enough for CBIR is not straightforward and may depend on how the reference standard is generated. For example, the intra- and inter-reader variability in our data show that some techniques for developing accurate reference standards for large databases, such as asking readers to view many images over extended periods of time, may be feasible, whereas other techniques, such as patching together multiple readers' ratings into a single similarity matrix, may prove challenging. To address these findings, future work includes developing CBIR systems that update their notions of similarity over time as individual readers use the system, and that can be customized by the user(s).

Acknowledgments

We would like to thank our participants for evaluating the images in this study, which included R. Brooke Jeffrey, MD, Rupesh Kalthia, MD, Dorcas Yao, MD, Ashwini Zenooz, MD, Adrian Albert, Jocelyn Barker, Tim Dorn, Daniel Golden, and Debra Willrett.

References

1. American Cancer Society, *Cancer Facts and Figures 2013*, American Cancer Society, Atlanta, GA (2013).
2. D. Marin et al., "Imaging approach for evaluation of focal liver lesions," *Clin. Gastroenterol. Hepatol.* **7**, 624–634 (2009).
3. I. R. Kamel, E. Liapi, and E. K. Fishman, "Liver and biliary system: evaluation by multidetector CT," *Radiol. Clin. North Am.* **43**, 977–997, vii (2005).
4. K. J. Mortele and P. R. Ros, "Cystic focal liver lesions in the adult: differential CT and MR imaging features," *RadioGraphics* **21**, 895–910 (2001).
5. H. Muller et al., "A review of content-based image retrieval systems in medical applications: clinical benefits and future directions," *Int. J. Med. Inf.* **73**, 1–23 (2004).
6. H. Muller et al., "Benefits of content-based visual data access in radiology," *RadioGraphics* **25**, 849–858 (2005).
7. H. Muller et al., "A review of content-based image retrieval systems in medical applications-clinical benefits and future directions," *Int. J. Med. Inf.* **73**(1), 1–23 (2003).
8. J. P. Eakins, "Towards intelligent image retrieval," *Pattern Recognit.* **35**, 3–14 (2002).
9. R. Datta et al., "Image retrieval: ideas, influences, and trends of the new age," *ACM Comput. Surv.* **40**, 5:1–5:60 (2008).
10. K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Comput. Med. Imaging Graphics* **31**, 198–211 (2007).
11. D. L. Akgul et al., "Content-based image retrieval in radiology: current status and future directions," *J. Digit. Imaging* **24**(2), 208–222 (2011).
12. C. Muramatsu et al., "Investigation of psychophysical similarity measures for selection of similar images in the diagnosis of clustered microcalcifications on mammograms," *Med. Phys.* **35**, 5695–5702 (2008).
13. B. M. Mehtre, M. S. Kankanhalli, and W. F. Lee, "Shape measures for content based image retrieval: a comparison," *Inform. Process. Manag.* **33**(3), 319–337 (1997).
14. J. Xu et al., "A comprehensive descriptor of shape: method and application to content-based retrieval of similar appearing lesions in medical images," *J Digit. Imaging* **25**, 121–128 (2012).
15. P. W. Huang and S. K. Dai, "Design of a two-stage content-based image retrieval system using texture similarity," *Inform. Process. Manag.* **40**(1), 81–96 (2004).
16. C. Lin, R. Chen, and Y. Chan, "A smart content-based image retrieval system based on color and texture feature," *Image Vision Comput.* **27**, 658–665 (2009).
17. G. Giacinto and F. Roli, "Bayesian relevance feedback for content-based image retrieval," *Pattern Recognit.* **37**, 1499–1508 (2004).
18. G. Duan, J. Yang, and Y. Yang, "Content-based image retrieval research," *Phys. Procedia* **22**, 471–477 (2011).
19. R. Zhang and Z. Zhang, "BALAS: empirical Bayesian learning in the relevance feedback for image retrieval," *Image Vision Comput.* **24**, 211–223 (2006).
20. J. Peng, B. Bhanu, and S. Qing, "Probabilistic feature relevance learning for content-based image retrieval," *Comput. Vision Image Understanding* **75**, 150–164 (1999).
21. Y. Liu et al., "A survey of content-based image retrieval with high-level semantics," *Pattern Recognit.* **40**, 262–282 (2007).
22. E. A. Krupinski and K. S. Berbaum, "The medical image perception society update on key issues for image perception research," *Radiology* **253**, 230–233 (2009).
23. E. A. Krupinski, "The role of perception in imaging: past and future," *Semin. Nucl. Med.* **41**, 392–400 (2011).
24. D. J. Manning, A. Gale, and E. A. Krupinski, "Perception research in medical imaging," *Br. J. Radiol.* **78**, 683–685 (2005).
25. C. A. Beam et al., "The place of medical image perception in 21st-century health care," *J. Am. Coll. Radiol.* **3**, 409–412 (2006).
26. C. Muramatsu et al., "Experimental determination of subjective similarity for pairs of clustered microcalcifications on mammograms: observer study results," *Med. Phys.* **33**, 3460–3468 (2006).
27. R. Nakayama et al., "Evaluation of objective similarity measures for selecting similar images of mammographic lesions," *J Digit. Imaging* **24**, 75–85 (2011).

28. C. Muramatsu et al., "Investigation of psychophysical measure for evaluation of similar images for mammographic masses: preliminary results," *Med. Phys.* **32**, 2295–2304 (2005).
29. C. Muramatsu et al., "Determination of similarity measures for pairs of mass lesions on mammograms by use of BI-RADS lesion descriptors and image features," *Acad. Radiol.* **16**, 443–449 (2009).
30. Q. Li et al., "Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules," *Med. Phys.* **30**, 2584–2593 (2003).
31. A. W. M. Smeulders et al., "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 1349–1380 (2000).
32. H. Muller et al., "Performance evaluation in content-based image retrieval: overview and proposals," *Pattern Recognit. Lett.* **22**, 593–601 (2001).
33. J. Faruque et al., "A scalable reference standard of visual similarity for a content-based image retrieval system," *IEEE Healthcare Inf. Imaging Syst. Biol.* 158–165 (2011).
34. J. Faruque et al., "Modeling perceptual similarity measures in CT images of focal liver lesions," *J. Digit. Imaging* **26**(4), 714–720 (2013).
35. J. L. Fleiss and J. Cohen, "The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability," *Educ. Psychol. Meas.* **33**(3), 613–619 (1973).
36. J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics* **33**, 159–174 (1977).
37. H. L. Kundel and M. Polansky, "Measurement of observer agreement," *Radiology* **228**, 303–308 (2003).

Jessica Faruque is currently a postdoctoral fellow at the National Institutes of Health. She received her PhD in electrical engineering from Stanford University in 2014. She received her master's degree in electrical engineering from Stanford University, and bachelor's degrees in electrical engineering and mathematics from California Polytechnic State University, San Luis Obispo. Her research focuses on medical imaging, image processing, and machine learning.

Christopher F. Beaulieu is a professor of radiology and chief of musculoskeletal imaging at Stanford University. He received his MD and PhD degrees from the University of Washington in Seattle, Washington, USA, and was a radiology resident at Duke University in Durham, North Carolina. He completed a fellowship in abdominal imaging at Stanford University. His research focuses on image processing, imaging informatics, and computer-aided detection in radiology.

Jarrett Rosenberg is a research scientist and biostatistician in the Department of Radiology at Stanford Medical School. He has extensive experience in both industrial and academic settings on experimental studies of measurement.

Daniel L. Rubin is an assistant professor of radiology and medicine (Biomedical Informatics Research) at Stanford University. He is PI of two centers in the NCI Quantitative Imaging Network (QIN), chair of QIN Executive Committee, and chair of Informatics Committee of the ECOG-ACRIN cooperative group. His NIH-funded research program focuses on quantitative imaging and techniques to integrate these data and discover imaging phenotypes that can predict underlying biology, define disease subtypes, and personalize treatment.

Dorcas Yao is a Stanford University affiliated clinical associate professor at the VA Palo Alto Health Care System. She is board-certified in clinical informatics and radiology, with 15+ years of clinical experience. She is passionate about improving health care delivery, with extensive experience in healthcare information technology and change management. She is also an MBA candidate with a concentration in medical management and leads projects aimed at improving operation, quality, processes, and outcomes.

Sandy Napel received his BS in engineering sciences from SUNY Stony Brook (1974), and his MS (1976) and PhD (1981) degrees in electrical engineering from Stanford University. He is a professor of radiology and, by courtesy, of electrical engineering and medicine at Stanford University. He co-leads the Stanford Radiology 3D and Quantitative Imaging Lab and the Section on Integrative Biomedical Imaging Informatics, where he is developing techniques linking image features to molecular properties of disease.