Automated Retrieval of CT Images of Liver Lesions on the Basis of Image Similarity: Method and Preliminary Results[1]

Sandy A. Napel, PhD
Christopher F. Beaulieu, PhD, MD
Cesar Rodriguez, MD
Jingyu Cui, MS
Jiajing Xu, MS
Ankit Gupta, BS
Daniel Korenblum, MS
Hayit Greenspan, PhD
Yongjun Ma, PhD
Daniel L. Rubin, MD, MS

Purpose: To develop a system to facilitate the retrieval of radiologic images that contain similar-appearing lesions and to perform a preliminary evaluation of this system with a database of computed tomographic (CT) images of the liver and an external standard of image similarity.

Materials and Methods: Institutional review board approval was obtained for retrospective analysis of deidentified patient images. Thereafter, 30 portal venous phase CT images of the liver exhibiting one of three types of liver lesions (13 cysts, seven hemangiomas, 10 metastases) were selected. A radiologist used a controlled lexicon and a tool developed for complete and standardized description of lesions to identify and annotate each lesion with semantic features. In addition, this software automatically computed image features on the basis of image texture and boundary sharpness. Semantic and computer-generated features were weighted and combined into a feature vector representing each image. An independent reference standard was created for pairwise image similarity. This was used in a leave-one-out cross-validation to train weights that optimized the rankings of images in the database in terms of similarity to query images. Performance was evaluated by using precision-recall curves and normalized discounted cumulative gain (NDCG), a common measure for the usefulness of information retrieval.

Results: When used individually, groups of semantic, texture, and boundary features resulted in various levels of performance in retrieving relevant lesions. However, combining all features produced the best overall results. Mean precision was greater than 90% at all values of recall, and mean, best, and worst case retrieval accuracy was greater than 95%, 100%, and greater than 78%, respectively, with NDCG.

Conclusion: Preliminary assessment of this approach shows excellent retrieval results for three types of liver lesions visible on portal venous CT images, warranting continued development and validation in a larger and more comprehensive database.

[1] From the Department of Radiology, Stanford University School of Medicine, James H. Clark Center S323, 318 Campus Dr, Stanford, CA 94305-5450 (S.A.N., C.F.B., C.R., D.L.R.); Departments of Electrical Engineering (J.C., J.X.) and Computer Science (A.G., D.K.), Stanford University, Stanford, Calif; Department of Computer Science, Tel Aviv University, Tel Aviv, Israel (H.G.); and Department of Computer Science and Information Engineering, Tianjin University of Science and Technology, Tianjin, China (Y.M.). Received September 13, 2009; revision requested November 4; revision received November 20; accepted December 9; final version accepted January 13, 2010. **Address correspondence to** S.A.N. (e-mail: *snapel@stanford.edu*).

Diagnostic radiologists are confronted with the challenge of efficiently and accurately interpreting cross-sectional studies that often contain thousands of images (1). Currently, this is largely an unassisted process, and a reader's accuracy is established through training and experience. Understanding how imaging features correlate with the underlying disease is central to radiologic training (2), and a considerable amount of radiology literature has been devoted to these correlations (3). Even so, there is substantial variation in interpretation between radiologists (4–6), and accuracy varies widely (7), a problem that is compounded as the number of images increases. Thus, there is an opportunity to improve diagnostic decision making by enabling radiologists to search databases of radiologic images and reports for cases that are similar in terms of shared imaging features to the cases on which they are working. However, at present, there is no systematic link between the actual pixel data in images and the descriptions of them in radiology reports and, as a result, the growing vast repositories of clinical imaging data cannot be searched effectively for similar images on the basis of descriptions of image features.

The goal of our study was to develop a content-based image retrieval system to facilitate the retrieval of radiologic images that contain similar-appearing lesions and to perform a preliminary evaluation of this system by using a database of computed tomographic (CT) images of the liver and an external standard of image similarity. Our system creates a vector of computer-accessible features (hereafter, feature vector) for lesions seen during imaging examinations and computes measures of similarity between feature vectors, establishing similarity between the corresponding images. The vectors contain detailed information about lesions, including *(a)* feature descriptors coded by radiologists using RadLex® (8), a comprehensive controlled terminology developed by the Radiological Society of North America, and *(b)* computer-generated features of pixels characterizing the interior texture of the lesion and the sharpness of its boundary. While diagnosis of liver lesions on the basis of CT results is a common and important clinical problem (9–16), our tools and algorithms are general so that they may be readily adapted to other modalities and diagnostic scenarios.

Advance in Knowledge

- Combinations of radiologist's descriptions with a controlled vocabulary and computer-derived features of lesion image texture and boundary sharpness can be used to retrieve similar images from an annotated image database.

Implication for Patient Care

- The ability to compare images with those obtained in other patients has the potential to provide real-time decision support to practicing radiologists by showing them similar images with associated diagnoses and, where available, responses to various therapies and outcomes.

Materials and Methods

Software Tools and Components

Our approach requires that lesions on CT images be defined by a region of interest drawn manually or automatically. In our current implementation, lesions are identified and circumscribed with an open-source medical image viewing application (OsiriX; *http://www.osirix-viewer.com/*) (17). Features are ascribed to each lesion in several ways, as described later in this article, and are entered into a database, making them available for combination with other features from the same lesion when computing similarity to other lesions.

Semantic Features

We developed a plug-in for the aforementioned medical image viewing software and called it the image Physician Annotation Device (hereafter, annotation device) (18). This plug-in provides an efficient and thorough means of capturing the semantic terms radiologists use to describe lesions (hereafter, semantic annotation). This plug-in also provides a structured data entry template that enables users to describe the features of lesions according to 12 categories comprising 161 possible descriptors (Table) selected from the RadLex controlled terminology (8) and augmented for this study with additional terms after consultation with an abdominal imaging radiologist. Once all entries have been made and validated against the controlled terminology of the annotation device, the annotations are saved in a file compliant with the Annotation and Image Mark-up standard (19,20), established by the National Cancer Institute's Cancer Bioinformatics Grid.

Computer-generated Features

Identified lesions can also be characterized with image processing methods. Some features may be more quantitative versions of semantic features, while others may represent features that are not directly appreciated by human observers but nonetheless may have value in discriminating among lesion types (21). Examples include the internal spatial distribution of pixel intensities (texture features) and analyses of the boundary between a given lesion and the surrounding tissue (boundary features).

Published online before print
10.1148/radiol.10091694

Radiology 2010; 256:243–252

Abbreviation:
NDCG = normalized discounted cumulative gain

Author contributions:
Guarantors of integrity of entire study, S.A.N., C.F.B., H.G., D.L.R.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; literature research, S.A.N., C.F.B., H.G., Y.M.; clinical studies, C.F.B., H.G.; statistical analysis, S.A.N., J.C., J.X., H.G., Y.M.; and manuscript editing, S.A.N., C.F.B., J.C., J.X., H.G., D.L.R.

Funding:
This research was supported by the National Institutes of Health (grant CA72023).

Authors stated no financial relationship to disclose.

Complete List of Available Lesion Descriptors by Category

Category of Lesion Description	Valid Selections
Overall lesion shape	Amorphous, asymmetrically shaped, beaded, curved, geographic, irregularly shaped*, linear, lobular*, nodular, macronodular, micronodular, ovoid*, pedunculated, platelike, polygonal, polypoid, round*, spoke-wheel, straightened, symmetrically shaped, wedge shaped, square, rectangular
Lesion margin and contours	Circumscribed margin*, irregular margin*, lobulated margin, obscured margin, poorly defined margin*, smooth margin*, spiculated margin, mixed margin, cluster of grapes
Lesion rim or capsule	Circumferential rim, incomplete rim, absent rim*, thick rim, thin rim, smooth rim, irregular rim, nodular rim, uniform rim, nonuniform rim
Lesion focality	Clustered, coalescent, diffuse, focal, multifocal, patchy, scattered, confluent, solitary lesion*, multiple lesions (2–5 lesions)*, multiple lesions (6–10 lesions)*, multiple lesions (>10 lesions)*, satellite lesion present, satellite lesions present
Lesion attenuation	
Overall	Fat density, hyperdense*, hypodense*, isodense, high signal intensity, low signal intensity, water density, soft-tissue density*, sparing of fat deposition, mildly dense, moderately dense, markedly dense, mixed density, mildly unrestricted diffusion, moderately unrestricted diffusion, markedly unrestricted diffusion, restricted diffusion, unrestricted diffusion
Uniformity within lesion	Homogeneous*, heterogeneous*, mixed
Lesion enhancement	
Overall within lesion	Enhancing*, isoenhancing, nonenhancing*, hypervascular, avascular
Spatial pattern	Circumferential enhancement, heterogeneous enhancement, homogeneous enhancement*, mosaic enhancement, mottled enhancement, peripheral continuous rim enhancement, peripheral discontinuous rim enhancement, peripheral continuous nodular enhancement, peripheral discontinuous nodular enhancement*, target enhancement, reticular enhancement, homogeneous enhancement except septa, linear gradient enhancement, radial gradient enhancement
Temporal features	Centripetal fill-in*, central fill-in, homogeneous retention*, homogeneous fade, homogeneous washout, mixed pattern, peripheral fill-in, peripheral retention, peripheral fade, peripheral washout
Lesion substance, other features	Calcification, hemorrhage, ischemia, necrosis, scar, internal nodules*, central scar, enhancing central scar, nonenhancing central scar, steatosis, vascular nidus, visible internal vessels
Lesion effect on liver	Abuts capsule of liver, bulges capsule of liver, retracts capsule of liver, invades into or along capsule of liver, invades through capsule of liver, abuts hepatic ducts, displaces hepatic ducts, retracts hepatic ducts, distorts hepatic ducts, obstructs hepatic ducts, infiltrates wall of hepatic ducts, infiltrates wall into lumen of hepatic ducts, abuts hepatic vein, displaces hepatic vein, retracts hepatic vein, distorts hepatic vein, obstructs hepatic vein, infiltrates wall of hepatic vein, infiltrates wall into lumen of hepatic vein, abuts portal vein, displaces portal vein, retracts portal vein, distorts portal vein, obstructs portal vein, infiltrates wall of portal vein, infiltrates wall into lumen of portal vein, abuts hepatic artery, displaces hepatic artery, retracts hepatic artery, distorts hepatic artery, obstructs hepatic artery, infiltrates wall hepatic artery, infiltrates wall into lumen of hepatic artery
Miscellaneous lesion findings	Feeding arteries, draining veins, perilesional vessels, normal perilesional tissue*, perilesional steatosis, perilesional sparing of focal fat, perilesional calcification, hemorrhagic perilesional tissue, perilesional perfusion alteration, hyperemic perilesional tissue

* Selected during this study as descriptive of the 30 CT images of liver lesions in our database.

While there are many types of computer-generated features in the literature, we chose to focus on these features on the basis of results reported for other content-based image retrieval applications (22–24). We note that particular features that turn out to be unimportant for a particular application will receive low weights during the learning process. This is discussed further in the Similarity Measure section.

Texture features.—For each lesion, we compute multiple features on the basis of pixels within the lesion region of interest. For example, for gray-level histogram-based analysis, 14 features—including the histogram itself (25), the low-frequency coefficients of its three-level Haar wavelet transform (26), the abscissa of its peak, and its variance—are computed. For analysis of Gabor features (27), the mean of the energy in the frequency domain over four scales and eight orientations in each of 32 bins is computed. This yields a total of 46 features.

Boundary features.—The boundary between a liver lesion and the surrounding normal liver contains critical diagnostic information; sharpness of the boundary generally reflects tumor aggressiveness and should help readers differentiate between lesions (11). We developed an automated method with which to characterize boundary sharpness that starts by enlarging the region of interest by using morphologic operations to include a margin of normal liver. Thereafter, attenuation values are bilinearly interpolated along radial line segments that are automatically drawn at many angles from the center outward to the dilated boundary. We then fit a

sigmoid function to each intensity profile by using the Levenberg-Marquardt algorithm (28) and use both the difference in intensity between the lesion and the surrounding liver and the sharpness of the margin to characterize each profile. Finally, we average each of these values across all profiles to generate two parameters that characterize mean boundary sharpness.

For each lesion, the combined output of the tools and software components described previously is a feature vector whose elements are made up of the semantic annotations and the computer-derived features. In our current implementation, each vector contains a total of 209 features: one hundred sixty-one semantic features, represented as binary values (present or absent); 46 computer-generated texture features (14 histogram features, 32 Gabor features); and two boundary features.

Similarity Measure

We defined the similarity of a pair of lesions as the inverse of a weighted sum of differences between corresponding elements of the respective feature vectors that describe them. We computed similarity by using semantic features alone, boundary features alone, histogram and Gabor features alone (in combination, they are known as texture features), and all four features in combination. Because we had a small data set compared with the number of individual elements in each feature vector, to avoid overfitting, we maintained equal weights for each member of each of the four groups. The first two types of similarity computations (which did not combine feature groups) were straightforward. For the latter two types of similarity computations, wherein feature groups were combined, we learned the weights from the data by using a modified version of a machine learning method known as adaptive boosting (AdaBoost) (29), which is described in the Training and Evaluation section.

Preliminary Evaluation

We designed our evaluation to assess the ability of our software tools, described previously, to assist us in ranking CT images of liver lesions stored in a database in order of visual similarity to a query lesion. We will now describe the lesion database, the similarity reference standard, our evaluation measures, and our methods for training and evaluation.

Lesion Database

We received institutional review board approval for retrospective analysis of deidentified patient images. Thereafter, we selected 30 portal venous phase CT images in which liver lesions had been identified from our clinical picture archiving and communication system. The 30 lesions (13 cysts, 10 metastases, and seven hemangiomas, according to the radiology reports) were in 15 patients (eight men, seven women; mean age, 56 years; age range, 39–88 years). We selected lesions of each type that radiologists would consider typical. Cysts were nonenhancing water-attenuation circumscribed lesions. Hemangiomas showed typical features of discontinuous nodular peripheral enhancement, with fill-in on delayed images. Metastases were hypoattenuating, had soft-tissue attenuation, enhanced homogeneously with contrast material administration, and had less well-defined margins than cysts. Images were acquired between February 2007 and August 2008, and the following parameters were used: 120 kVp, 140–400 mAs, and 2.5–5.0-mm section thickness. These types of lesions are common and span a range of appearances. A radiologist (C.F.B., 15 years of abdominal CT experience) used the aforementioned medical image viewing software to circumscribe each lesion boundary and used the annotation device plug-in to choose from among the 161 semantic descriptors to annotate each lesion; descriptors that were selected marked the presence of the feature, and those not selected indicated absence.

Reference Standard

We created a separate reference standard of image similarity for the 30 CT images of liver lesions described previously to enable us to evaluate image retrieval by using the semantic and computer-generated features alone and in combination. Two radiologists (C.F.B., D.L.R.; 15 and 10 years of experience in abdominal imaging, respectively) viewed each pair of images twice and reached a consensus opinion on a similarity measure for the pair (3, very similar; 2, somewhat similar; 1, not similar) by addressing similarity of texture, boundary shape, and sharpness. They did not consider size or location within the liver, nor did they consider any clinical data that might have implied a specific lesion type. They strived to base their evaluation purely on image appearance and to not classify the lesions by diagnosis. Thus, with this reference standard, a perfect retrieval system would return a sequence of images with similarities monotonically decreasing from 3 to 1.

Evaluation Measures

We used two measures to evaluate performance. Precision recall (30), which is commonly used in the content-based retrieval community, plots precision (the number of similar images retrieved divided by the total number of images retrieved) versus recall (the number of similar images retrieved divided by the total number of similar images in the database). Similar to the receiver operating characteristic (31) analysis commonly used in radiology research, this type of evaluation requires binary truth (the retrieved image must be defined in the reference standard to be similar or dissimilar to the query image). Thus, we calculated the average similarity for the two viewings of each pair (which were always agreed to within one unit), and calculated a threshold for the result to define *similar* as having an average similarity score in the reference standard of 2.5 or greater.

We also used normalized discounted cumulative gain (NDCG) (32), which is a standard technique used to measure the effectiveness of information retrieval algorithms when graded truth is available, as represented by our three-point similarity scale. We did not use the average of the two viewings of each pair, nor did we threshold the reference standard for this analysis. NDCG is used

to measure the usefulness (gain) on a scale of 0 to 1 of K retrieved lesions on the basis of their positions in the ranked list compared with their similarity to the query lesion according to a separate reference standard. The accumulated gain is evaluated with the weight of each retrieved lesion discounted at lower ranks. Thus, for a given K, higher NDCG(K) means more lesions similar to the query image are ranked ahead of dissimilar ones, with NDCG(K) equal to 1 implying perfect retrieval of K images.

Training and Evaluation

For the evaluation of two individual feature categories (semantic, boundary), there is no relative weighting of features; thus, training is not required. We withheld each image from the database and ranked the remaining 29 images according to our similarity measure. For each query image, we computed NDCG at each K and precision at each of 10 values of recall over all withheld images. For texture features, which are a combination of histogram and Gabor features, and for the combination of all feature categories, we used a modified version of adaptive boosting (AdaBoost) (29) in a leave-one-out validation framework to learn weights for each feature category, as follows: As before, we withheld one of the 30 images. We used the similarity reference standard described previously with the remaining 29 images and computed optimal values for the weights (feature weights) to be used in the similarity computation, as follows: We first initialized the feature weights and a weight for each of the 29 lesions (lesion weights) to unity. We then performed the following 200 times: We randomly sampled a subset of the 29 lesions, with random size D ($D \geq 1$ and $D \leq 29$), with probabilities proportional to the lesion weights. For each feature group separately, we selected each of the D lesions in turn, ranked the remaining $D-1$ lesions in order of similarity, and used the similarity reference standard to compute performance, defined as the average over all D lesions of the average NDCG(K) over all K = 1 to $D-1$. We selected the feature group in which performance was best, increased the relative weight for that feature, and decreased the relative weight for the D lesions used in this iteration by an amount proportional to this performance. In this way, better-performing features were given more weight in the similarity calculation, and better-performing lesions (easier lesions) were less likely to be sampled again. Thus, for every withheld lesion, an optimal set of weights was computed without benefit of the withheld one. These weights were then used in the similarity calculation to rank the remaining 29 lesions which, combined with the similarity standard for the 29 lesions, generated a single NDCG curve. Finally, for each of the four analyses (semantic, boundary, texture, combined), we computed the mean and standard deviation of NDCG over all 30 withheld lesions at each K = 1, 29, and of precision at each of 10 values of recall.

Results

Semantic annotation required approximately 3 minutes per lesion. The number of descriptors chosen per lesion ranged from eight to 11 (mean, 9.2 descriptors ± 1.2 [standard deviation]) and, in total, 26 unique descriptors were selected to indicate features present in these 30 lesions.

The similarity reference standard had the following distribution of values: Of the 870 pairs, 219 (25%) were assigned a similarity score of 3 (similar), 141 (16%) were assigned a similarity score of 2 (somewhat similar), and 510 (59%) were assigned a similarity score of 1 (not similar).

Figure 1 shows precision-recall plots obtained by using texture, boundary, and semantic features alone and in combination. As ideal precision-recall plots hug the line indicating a precision of 1.0 for all values of recall, combining all features produced the best results, with average precision values of more than 90% (meaning, on average, nine of 10 images were rated to have similarity to the reference standard of 2.5 or greater) for all values of recall.

Figure 2 shows the NDCG results obtained by using texture, boundary, and semantic features alone and in combination. Again, combining all features yielded the best overall results, with mean, best, and worst case retrieval accuracy of greater than 95%, 100%, and greater than 78%, respectively, for all values of K, yielding what would be considered excellent information retrieval performance.

While the precision-recall and NDCG plots show performance over all 30 possible query images, Figure 3 exemplifies retrieval results by using one of each type of lesion. Perfect retrieval would result in a ranked order of images with monotonically decreasing reference standard similarities. None of the three examples shown yielded a perfect ranking, but all three yielded a reasonable one. Images in the reference standard that were judged to be dissimilar to the query image (similarity of 1) never appeared ahead of images judged to be similar (similarity of 3) to the query image.

Discussion

We have developed a content-based retrieval system that incorporates semantic features observed by radiologists, as well as features computationally extracted from the images themselves, and shown it to be capable of yielding excellent retrieval results. Use of semantic information associated with images is not new. In fact, figure legends from hundreds of peer-reviewed journals can be searched with the Goldminer tool (33), and text-based searches are the basis of the Google Images search engine and many other Web applications. However, the semantic information associated with the vast majority of radiologic images exists only in radiology reports. While these reports can be used for image retrieval (34), this approach is problematic for the following reasons: First, reports generally refer to studies consisting of many images. Second, descriptions of lesions are not mapped to specific image pixels. Third, radiologists may use different terms to describe the same observation (35,36). Fourth, complete description of lesions is not enforced. Fifth, negation, often appearing in reports, can be confusing

Figure 1

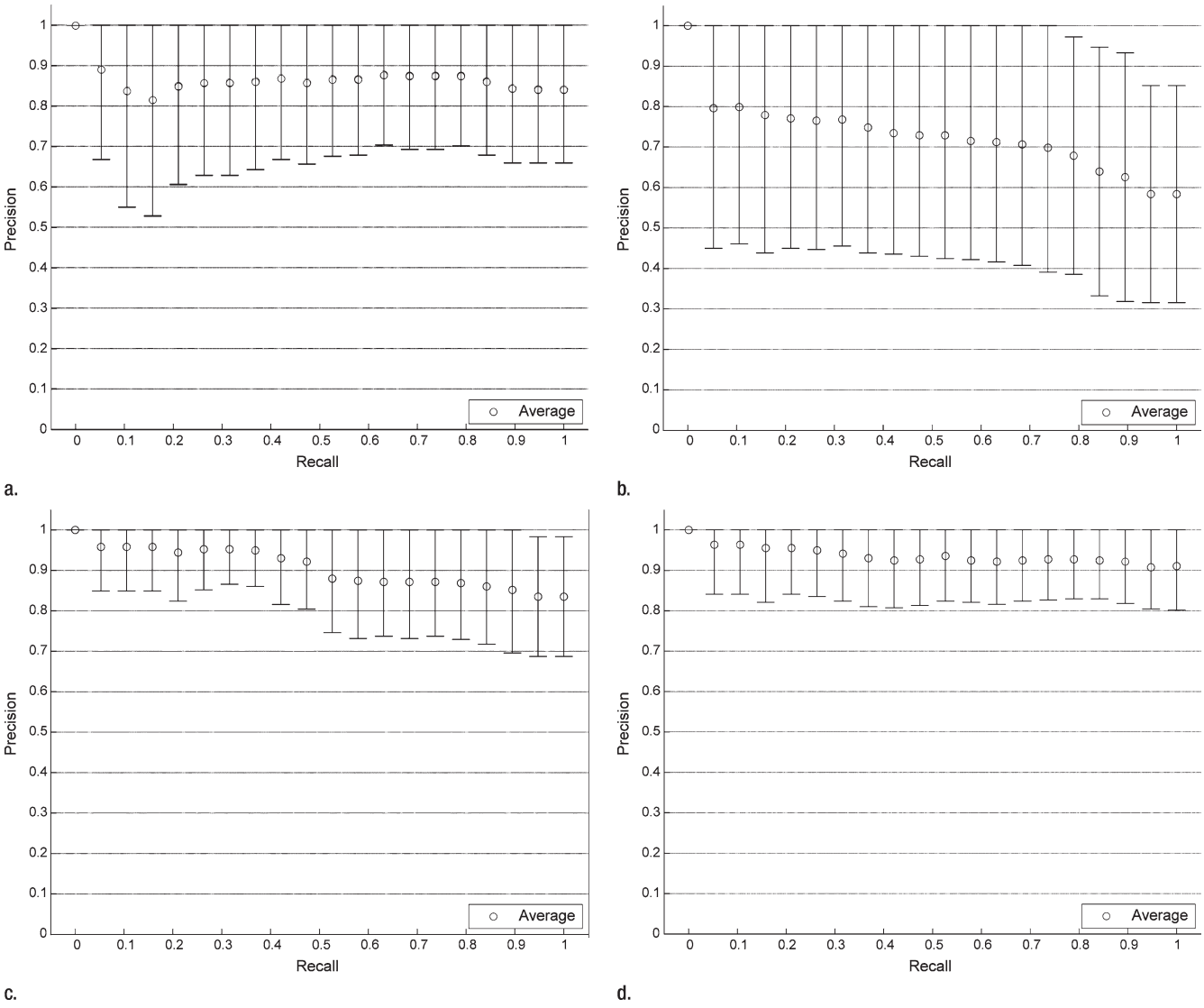

Figure 1: **(a–d)** Precision-recall plots for different types of image features show average performance in a set of 30 images containing cysts, metastases, and hemangiomas. Error bars are less than or equal to one standard deviation. **(a)** Texture (combined histogram and Gabor) features, **(b)** boundary features, **(c)** semantic features, and **(d)** all features combined.

(37). As a result, little semantic content from the reporting process can be machine accessible or associated with specific images. Our structured approach to semantic annotation addresses all of these issues, providing consistent descriptors mapped to specific regions of specific images. With our application, an average of only 9.2 features per lesion was selected relatively quickly; however, it may be too cumbersome to use in other clinical scenarios. Nonetheless, our study shows the utility of complete description with a controlled vocabulary, encouraging the development of more efficient systems in the future.

We note that of the 161 unique descriptors offered by the annotation device for this clinical application, only 26 descriptors were ultimately selected. This is probably because of the limited number of lesion types in this study. We expect that more of the available terms will be used as our database grows to include a broader range of lesion types.

The use of algorithmic feature extraction from images has also been applied to content-based image retrieval both outside (22) and inside (30,34,38,39) radiologic applications. In gray-scale radiologic images, texture features that use histograms and wavelets have proved powerful in feature extraction (40–43) and may be used to capture important features that are not visually apparent

Figure 2

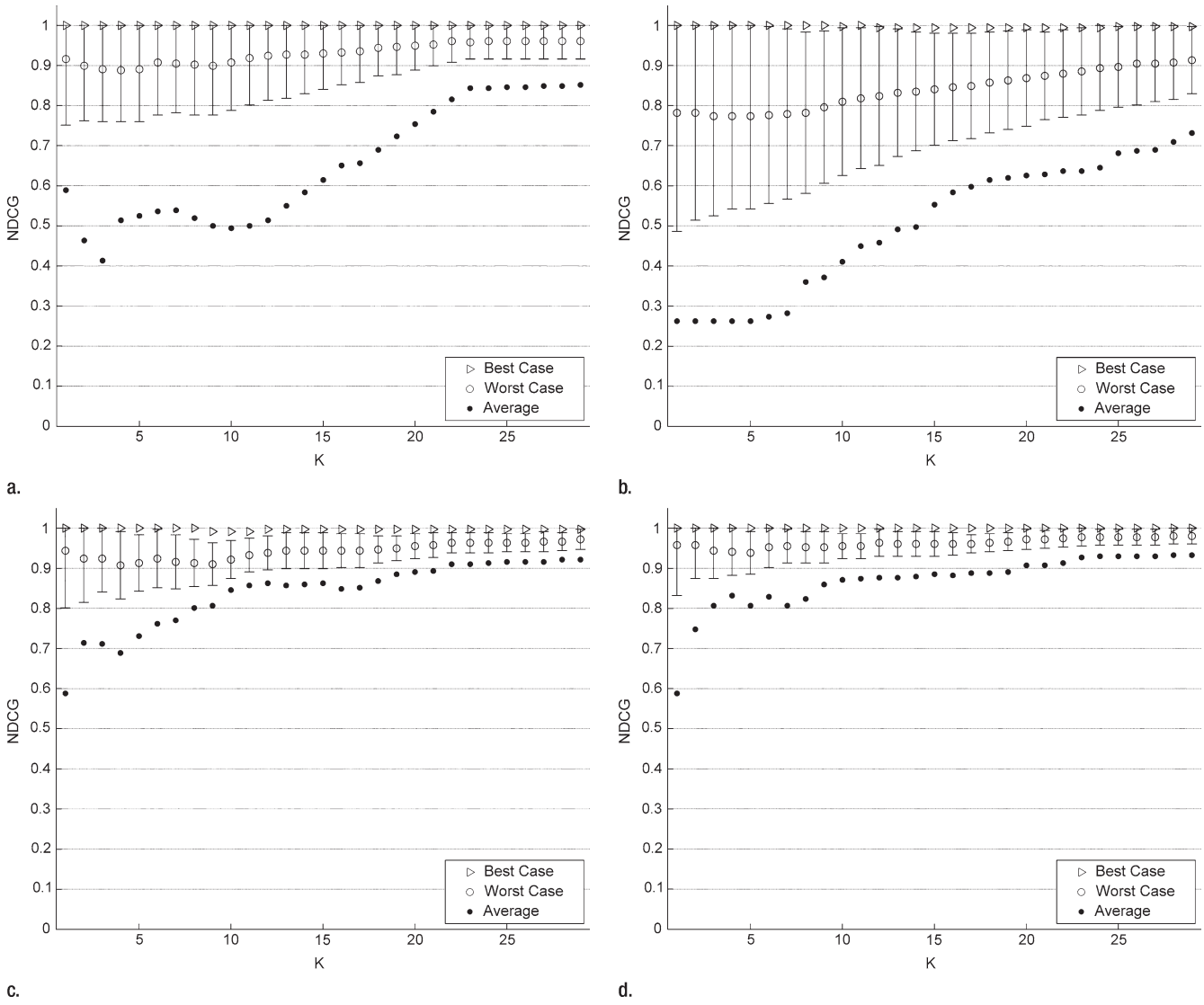

Figure 2: **(a–d)** NDCG(K) plots, where K is the number of images retrieved, for different types of image features show average, best, and worst case performance in a set of 30 images containing cysts, metastases, and hemangiomas. Error bars are less than or equal to one standard deviation. **(a)** Texture (combined histogram and Gabor) features, **(b)** boundary features, **(c)** semantic features, and **(d)** all features combined.

(21). Most often, these methods have been applied to entire images, not specific regions within them; however, there have been good results in the specific area of lung nodule classification and retrieval (44). We focused on liver lesion retrieval and showed that inclusion of carefully obtained semantic features in addition to computer-generated features is beneficial.

In our study, we evaluated performance of our system with two measures. Precision-recall (30) calculations require binary truth; a retrieved image is either similar or dissimilar to the query image. When truth is graded (as in our three-point similarity scale), this requires the definition of an arbitrary threshold, the choice of which may affect the results. In addition, precision-recall calculations do not take into account the retrieved image order. NDCG accounts for ordering and graded truth; thus, it may be a more appropriate performance measure for our system. While both measures showed that our system performs well, there was less variance in the NDCG results.

Our study had several limitations, including the use of a small validation data set (n = 30) involving only three types of relatively distinctive lesions and the fact that more than one lesion was used per patient. Besides the obvious issues with generalizability, the small size of our data set required us to limit the

Figure 3

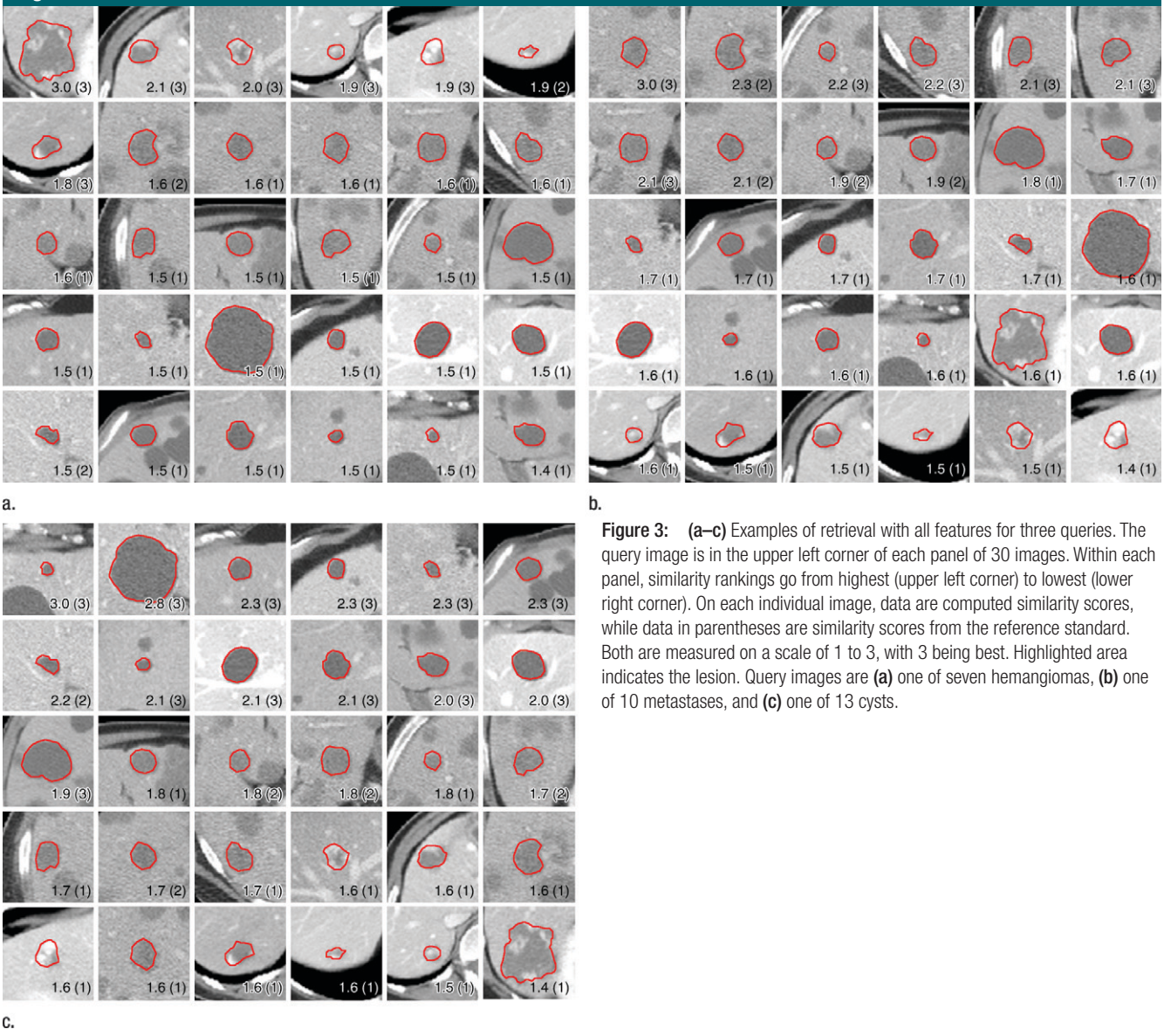

Figure 3: **(a–c)** Examples of retrieval with all features for three queries. The query image is in the upper left corner of each panel of 30 images. Within each panel, similarity rankings go from highest (upper left corner) to lowest (lower right corner). On each individual image, data are computed similarity scores, while data in parentheses are similarity scores from the reference standard. Both are measured on a scale of 1 to 3, with 3 being best. Highlighted area indicates the lesion. Query images are **(a)** one of seven hemangiomas, **(b)** one of 10 metastases, and **(c)** one of 13 cysts.

number of feature-combining weights to avoid overfitting during the training phase. The use of more weights would allow us to take advantage of the high dimensionality of the feature space and could lead to better understanding of the important features and improved performance. We note that although the machine-learning approach we adopted (adaptive boosting) performed well, it may not perform as well as others in this small data set with fewer patients than selected lesions. Also, as there was no objective way to create a reference standard for image similarity, our reference standard served as only a reasonable approximation to truth.

Another limitation was that our system relies on identification and segmentation of lesions, which currently is performed by humans and therefore may be subject to inter- and intraobserver variability. Future developments of automated detection and segmentation techniques can be incorporated and should be beneficial. Because lesion boundary sharpness is a function of section thickness, in-plane pixel size (partial volume), and lesion infiltration, more consistency in these reconstruction parameters may improve its performance. Finally, we recognize that there are a wide variety of appearances of benign and malignant lesions and that multiphasic imaging is central to current clinical diagnosis. While we are not proposing that the current preliminary results encompass the full field of liver CT, our results do suggest that it is feasible to combine human- and machine-annotated image features to retrieve images that are similar

to a query image from a database of CT images of liver lesions, thereby encouraging continued development and evaluation with a larger and more comprehensive database.

In conclusion, content-based retrieval of images containing similar-appearing lesions is practical; preliminary assessment of our approach shows reasonable retrieval results compared with an independently constructed pairwise visual similarity standard for three types of liver lesions visible on portal venous CT images. The technology we have developed is general and can be easily adapted to other anatomic and diagnostic scenarios in which CT and other imaging modalities are used. Ultimately, our approach could provide real-time decision support to practicing radiologists by showing them similar images with associated diagnoses and, where available, responses to various therapies and outcomes.

Acknowledgments: We are grateful for assistance from Claude B. Sirlin, MD, and Cynthia Santillan, MD, in defining additional semantic features of liver lesions not yet included in the RadLex lexicon.

References

1. Rubin GD. Data explosion: the challenge of multidetector-row CT. Eur J Radiol 2000; 36(2):74–80.
2. Farzanegan F. Keep AFIP. J Am Coll Radiol 2006;3(12):961.
3. Kreel L, Arnold MM, Lo YF. Radiological-pathological correlation of mass lesions in the liver. Australas Radiol 1991;35(3):225–232.
4. Armato SG 3rd, McNitt-Gray MF, Reeves AP, et al. The Lung Image Database Consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. Acad Radiol 2007;14(11): 1409–1421.
5. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst 2004;96(24):1840–1850.
6. Hillman BJ, Hessel SJ, Swensson RG, Herman PG. Improving diagnostic accuracy: a comparison of interactive and Delphi consultations. Invest Radiol 1977;12(2):112–115.
7. Robinson PJ. Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image. Br J Radiol 1997;70(839): 1085–1098.
8. Langlotz CP. RadLex: a new method for indexing online educational materials. RadioGraphics 2006;26(6):1595–1597.
9. Brenner DJ, Hall EJ. Computed tomography: an increasing source of radiation exposure. N Engl J Med 2007;357(22):2277–2284.
10. Kamel IR, Liapi E, Fishman EK. Liver and biliary system: evaluation by multidetector CT. Radiol Clin North Am 2005;43(6): 977–997, vii.
11. Mortelé KJ, Ros PR. Cystic focal liver lesions in the adult: differential CT and MR imaging features. RadioGraphics 2001;21(4): 895–910.
12. El-Serag HB, Marrero JA, Rudolph L, Reddy KR. Diagnosis and treatment of hepatocellular carcinoma. Gastroenterology 2008;134(6):1752–1763.
13. Llovet JM, Burroughs A, Bruix J. Hepatocellular carcinoma. Lancet 2003;362(9399): 1907–1917.
14. Marin D, Furlan A, Federle MP, Midiri M, Brancatelli G. Imaging approach for evaluation of focal liver lesions. Clin Gastroenterol Hepatol 2009;7(6):624–634.
15. Blachar A, Federle MP, Ferris JV, et al. Radiologists' performance in the diagnosis of liver tumors with central scars by using specific CT criteria. Radiology 2002;223(2): 532–539.
16. Seltzer SE, Getty DJ, Pickett RM, et al. Multimodality diagnosis of liver tumors: feature analysis with CT, liver-specific and contrast-enhanced MR, and a computer model. Acad Radiol 2002;9(3):256–269.
17. Rosset A, Spadola L, Ratib O. OsiriX: an open-source software for navigating in multidimensional DICOM images. J Digit Imaging 2004;17(3):205–216.
18. Rubin DL, Rodriguez C, Shah P, Beaulieu C. iPad: semantic annotation and markup of radiological images. AMIA Annu Symp Proc 2008:626–630.
19. Channin DS, Mongkolwat P, Kleper V, Sepukar K, Rubin DL. The caBIG Annotation and Image Markup Project. J Digit Imaging 2010;23(2):217–225.
20. Rubin DL, Mongkolwat P, Kleper V, Supekar K, Channin DS. Annotation and image markup: accessing and interoperating with the semantic content in medical imaging. IEEE Intell Syst 2009;24(1):57–65.
21. Brown R, Zlatescu M, Sijben A, et al. The use of magnetic resonance imaging to noninvasively detect genetic signatures in oligodendroglioma. Clin Cancer Res 2008; 14(8):2357–2362.
22. Datta R, Joshi D, Li J, Wang JZ. Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 2008;40(2): 1–60.
23. Mirmedhi M, Xie X, Suri J, eds. Handbook of texture analysis. London, England: Imperial College Press, 2008.
24. Li J, Wang JZ. Real-time computerized annotation of pictures. IEEE Trans Pattern Anal Mach Intell 2008;30(6):985–1002.
25. Bilello M, Gokturk SB, Desser T, Napel S, Jeffrey RB Jr, Beaulieu CF. Automatic detection and classification of hypodense hepatic lesions on contrast-enhanced venous-phase CT. Med Phys 2004;31(9):2584–2593.
26. Strela V, Heller PN, Strang G, Topiwala P, Heil C. The application of multiwavelet filterbanks to image processing. IEEE Trans Image Process 1999;8(4):548–563.
27. Zhao CG, Cheng HY, Huo YL, Zhuang TG. Liver CT-image retrieval based on Gabor texture. In: IEMBS '04. 26th Annual International Conference of the IEEE. San Francisco, Calif; Engineering in Medicine and Biology Society, 2004:1491–1494.
28. Hagan MT, Menhaj MB. Training feedforward networks with the Marquardt algorithm. IEEE Trans Neural Netw 1994;5(6): 989–993.
29. Freund Y, Schapire RE. A short introduction to boosting. J Jpn Soc Artif Intell 1999;14(5):771–780.
30. Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications: clinical benefits and future directions. Int J Med Inform 2004;73(1):1–23.
31. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143(1):29–36.
32. Jarvelin K, Kekalainen J. Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 2002;20(4):422–446.
33. Kahn CE Jr, Thao C. GoldMiner: a radiology image search engine. AJR Am J Roentgenol 2007;188(6):1475–1478.
34. Hersh W, Müller H, Kalpathy-Cramer J. The ImageCLEFmed medical image retrieval task test collection. J Digit Imaging 2009; 22(6):648–655.
35. Sobel JL, Pearson ML, Gross K, et al. Information content and clarity of radiologists' reports for chest radiography. Acad Radiol 1996;3(9):709–717.

36. Stoutjesdijk MJ, Fütterer JJ, Boetes C, van Die LE, Jager G, Barentsz JO. Variability in the description of morphologic and contrast enhancement characteristics of breast lesions on magnetic resonance imaging. Invest Radiol 2005;40(6):355–362.

37. Lowe HJ, Antipov I, Hersh W, Smith CA. Towards knowledge-based retrieval of medical images: the role of semantic indexing, image content representation and knowledge-based retrieval. Proc AMIA Symp 1998: 882–886.

38. Greenspan H, Pinhas AT. Medical image categorization and retrieval for PACS using the GMM-KL framework. IEEE Trans Inf Technol Biomed 2007;11(2):190–202.

39. Image Retrieval in CLEF. ImageClef Web site. http://imageclef.org/. Accessed April 15, 2009.

40. Bovik AC, Clark M, Geisler WS. Multichannel texture analysis using localized spatial filters. IEEE Trans Pattern Anal Mach Intell 1990;12(1):55–73.

41. Mlsna PA, Sirakov NM. Intelligent shape feature extraction and indexing for efficient content-based medical image retrieval. 6th IEEE Southwest Symposium on Image Analysis and Interpretation, 2004. Piscataway, NJ: IEEE, 2004;172–176.

42. Viola PA, Jones MJ. Rapid object detection using a boosted cascade of simple features. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, 2001. Piscataway, NJ: IEEE, 2001; 511–518.

43. Wang JZ, Wiederhold G, Firschein O, Wei SX. Content-based image indexing and searching using Daubechies' wavelets. Int J Digit Libr 1998;1(4):311–328.

44. Lam MO, Disney T, Raicu DS, Furst J, Channin DS. BRISC: an open source pulmonary nodule image retrieval framework. J Digit Imaging 2007;20(suppl 1):63–71.