Experimental determination of subjective similarity for pairs of clustered microcalcifications on mammograms: Observer study results

Chisako Muramatsu, Qiang Li, Robert Schmidt, Kenji Suzuki, Junji Shiraishi, Gillian Newstead, and Kunio Doi
Department of Radiology, The University of Chicago, Chicago, Illinois 60637

(Received 9 March 2006; revised 1 June 2006; accepted for publication 13 July 2006; published 30 August 2006)

Presentation of images of lesions similar to that of an unknown lesion might be useful to radiologists in distinguishing between benign and malignant clustered microcalcifications on mammograms. Investigators have been developing computerized schemes to select similar images from large databases. However, whether selected images are really similar in appearance is not examined for most of the schemes. In order to retrieve images that are useful to radiologists, the selected images must be similar from radiologists' diagnostic points of view. Therefore, in this study, the data of radiologists' subjective similarity for pairs of clustered microcalcification images were obtained from a number of observers, and the intra- and inter-observer variations and the intergroup correlations were determined to investigate whether reliable similarity ratings by human observers can be determined. Nineteen images of clustered microcalcifications, each of which was paired with six other images, were selected for the observer study. Thus, subjective similarity ratings for 114 pairs of clustered microcalcifications were determined by each observer. Thirteen breast, ten general, and ten nonradiologists participated in the observer study; some of them completed the study multiple times. Although the intraobserver variations for the individual readings and the interobserver variations for pairs of observers were not small, the interobserver agreements were improved by taking the average of readings by the same observers. When the similarity ratings by a number of observers were averaged among the groups of breast, general, and nonradiologists, the mean differences of the ratings between the groups decreased, and good concordance correlations (0.846, 0.817, and 0.785) between the groups were obtained. The result indicates that reliable similarity ratings can be determined by use of this method, and the average similarity ratings by breast radiologists can be considered meaningful and useful for the development and evaluation of a computerized scheme for selection of similar images. [DOI: 10.1118/1.2266280]

Key words: mammograms, clustered microcalcifications, similar images, observer study

I. INTRODUCTION

Breast cancer is the second leading cause of cancer death and the most frequently diagnosed nonskin cancer in women in the United States. The American Cancer Society[1] estimates that 212 920 new invasive cancer and 61 980 new *in situ* breast cancer cases will be diagnosed in 2006. Although mammography is considered useful as an early detection tool, there are still false negative studies.[2–4] A number of studies[2,4–7] has reported that the computer-aided diagnosis (CAD), defined as a diagnosis made by a radiologist who takes into consideration a "second opinion" provided by a computer, may be useful in the detection of breast lesions on mammograms. In fact, commercial systems for aided detection of lesions on mammograms have been approved for clinical use by the Food and Drug Administration, and CAD is employed[8] at many clinical facilities in the U.S. However, detection is only part of the diagnostic task. Once detected, it can be difficult to distinguish between malignant and benign lesions on mammograms. Investigators have therefore been developing CAD schemes for characterization of detected lesions to help radiologists in reducing the "unnecessary" recall examinations and the number of biopsies of benign lesions. In most of these CAD,[9–13] computers provide radiologists the likelihood of malignancy in percentage format, without specific reasons as to why the likelihood is high or low. When a radiologist encounters an unknown lesion, we can assume that he/she tries to recall similar cases that he/she has experienced in clinical practice and/or learned in training courses or from textbooks. Therefore, to complement an estimated numerical likelihood, we believe that the presentation of images with known diagnoses similar to that of an unknown lesion would be helpful. In fact, image retrieval methods, such as keyword searching[14,15] and content-based or feature-based retrieval,[16–21] have been studied by investigators for the purpose of a diagnostic aid or teaching tool. Nonetheless, in most of these studies, whether retrieved images were really similar in appearance was not examined subjectively by radiologists.

In order for retrieved images to be really helpful in CAD, we believe that the images must be visually similar from radiologists' diagnostic point of view. Thus, our hypothesis is that reliable data on radiologists' impression of similarity for many types of image pairs would be useful for development of such CAD schemes and for the evaluation of the useful-

ness of image retrieval methods. Li *et al.*[22] conducted an observer study to determine subjective similarity for pairs of lung nodules in computed tomography (CT), which was then used for determination and evaluation of similarity measures. However, the appearance of lung nodules on CT is very different from that of microcalcifications on mammograms, and normal structures included in thoracic CT and mammograms are also very different. The determination of subjective similarity may be more difficult for pairs of clustered microcalcifications, because radiologists usually take into account the shapes and distribution of individual microcalcifications as well as the cluster as a whole. Image resolutions are different for CT (order of millimeter) and digital mammograms (typically 50 to 100 μm). Image presentation (reconstructed slice images versus projection images, respectively) is also different. A study was reported by Nishikawa *et al.*[23] in which they compared two methods for determination of similarity scores for pairs of clustered microcalcifications; however, the number of cases and the number of observers were limited. El-Naqa *et al.*[24] obtained subjective similarity scores for pairs of clustered microcalcifications based only on the spatial characteristics of the clusters. Images retrieved by use of such similarity scores, however, might not be helpful for diagnosis, because images would be "similar" only in terms of the cluster distribution. The purpose of this study is to measure and quantify radiologists' subjective similarity for pairs of images with clustered microcalcifications based on the overall impression for diagnosis and to investigate the variations within and between observers.

II. MATERIAL AND METHODS

A. Images of clustered microcalcifications used in this study

Images of clustered microcalcifications were obtained from a publicly available database, the Digital Database for Screening Mammography (DDSM),[25] developed by researchers at the University of South Florida and others. The DDSM includes images of 914 biopsy-proven cancer cases, 996 (141 not biopsy-proven) benign cases, and 695 normal cases collected from four facilities from 1988 to 1999. For each lesion identified, the rough outline of the lesion, the subtlety, and breast imaging reporting and data system (BI-RADS) description and assessment were included in the DDSM. For this study, 881 square regions (5 cm$\times$5 cm) of interest (ROIs), including 378 and 503 ROIs with malignant and benign clustered microcalcification lesions, respectively, were obtained. For all of the ROIs obtained, individual microcalcifications were identified based on the outlines of the lesions by an experienced technologist (H.N.) for computerized image analysis. The contrast and the density level for all of the ROIs were manually adjusted by a breast radiologist (R.A.S.) for optimal viewing.

B. Observer study for determination of similarity

To obtain subjective similarity for pairs of clustered microcalcifications, 19 sets of images (= ROIs) were prepared.

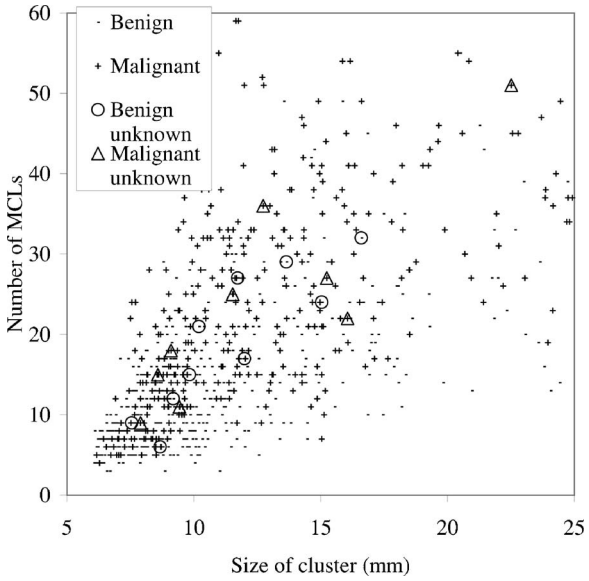

FIG. 1. Characteristics in effective diameter of the cluster and the number of microcalcifications in the cluster of all the images and ones used as "unknown" images.

In each set, one image was placed in the center ("unknown" image) with three images placed each on the right and left ("known" images) which were compared to the center image. Thus, six similarity ratings were obtained for a set, and the similarity ratings for a total of 114 image pairs (19×6) were determined. First, nine malignant and ten benign "unknown" images were selected by the breast radiologist (R.A.S.) to provide a variety of types of clustered microcalcifications. Figure 1 shows the characteristics in the effective diameter of the cluster and the number of microcalcifications in the cluster for "unknown" images used. For each unknown image, about 10 to 40 malignant and benign candidates for "known" images were automatically selected by use of image features, such as the number, the contrast, and the shape irregularity of microcalcifications and the size of the cluster. If these cases were selected randomly, most pairs would be dissimilar, and such data would not be useful. Therefore, to ensure six similarity ratings to be distributed in a wide range, final selections were made by the consensus of three coauthors (Q.L., K.D., and C.M.). Since some "known" images were used more than once, 113 different ROIs obtained from 101 patients were employed in these 114 pairs. All of the identifications in the DDSM for the 113 ROIs used in the study are listed in the Appendix. For the 113 ROIs, the number of identified microcalcifications in the lesion ranged from 5 to 68 with the average of 20. The effective diameter of the lesion ranged from 7 to 23 mm with the mean of 11.7 mm.

The order of 19 sets as well as the placement of six "known" images was randomized, and pathologies of lesions were not revealed to the observers. The images were displayed on a monochrome liquid crystal display monitor (ME511L/P4, 21.3 in., 2048×2560 pixels, 410 cd/m^2 luminance; Totoku Electric Co., Ltd.) in full resolution (zoomed mode) with the capability of unzooming. In zoomed mode, the size of each image was 3 cm$\times$3 cm, showing the entire

TABLE I. The number of observers participated in the observer study.

	Breast radiologists	General radiologists	Non-radiologists
Five readings	1	1	5
Two readings	8	0	3
One reading	4	9	2
Total	13	10	10

lesion. The observers were asked to mark their impression of similarity by clicking with a mouse on a continuous rating scale between 0 and 1, corresponding to two images that were not similar at all and almost identical, respectively. The data were quantified automatically in an observer interface program and stored electronically. The instructions to the observers were (1) Purpose: To obtain basic data for selecting similar images in CAD scheme to assist radiologists' interpretation of mammograms; (2) Cases: Nineteen unknown clustered microcalcifications (approximately equal number of malignant and benign) together with six similar or dissimilar clustered microcalcifications; (3) Similarity rating: Based on your overall impression for radiological diagnosis, use continuous rating scale with a line-checking method where two clustered microcalcifications are 0: not similar at all and 1: almost identical; (4) Rating: Each should be rated independently and consistently; and (5) Reading time: No time limit. At the beginning of the observer study, a training session with two "unknown" cases, i.e., ratings for 12 pairs of images, was provided for the observers to familiarize themselves with the rating method. During the training session, the observers can experience the range of similarity expected in the subsequent study so that they could scale their impression.

A total of 33 observers including 13 breast radiologists, 10 general radiologists (one resident), and 10 nonradiologists participated in the study. Some observers completed the study multiple times, and the numbers were summarized in Table I. The orders of cases and the placement of six images were randomized in each of repeated studies. The reproducibility within each observer and the agreement between two observers were assessed in terms of the concordance correlation coefficient.[26,27] The concordance correlation is a modification of the Pearson's correlation coefficient. Unlike the Pearson's correlation which is a measure of linear association, the concordance correlation takes into account the deviation of the best-fit line from the 45 deg line. There are other methods to assess agreement, such as the Bland-Altman method[28] and intraclass correlation coefficient.[29] Bland and Altman suggested that differences should be plotted against the mean, instead of one (rater) versus the other (rater). The Bland-Altman method can be used to detect whether the fixed and/or the proportional bias exists. With the Bland-Altman method, the limits of agreement are determined, in which most of the differences are expected to be found. However, it is difficult to interpret how well the agreement is, whereas with correlation coefficient, it is easier to understand that 1.0 and 0.0 correspond to a perfect agreement and no agreement, respectively. The intraclass correlation coefficient (ICC) is a fraction of the between-class variation to the total variation. When the variance between cases is much larger than the variance between observers within cases, ICC becomes high. The ICC for a pair of raters is very similar to the concordance correlation, especially when the variance of differences between the two is small or the number of cases is large.[30] The standard deviation of ratings for each pair of clustered microcalcifications was also determined to examine the intraobserver variability and interobserver variability for each group of observers. The similarity ratings were averaged first for each observer and then within the group of observers to determine intergroup correlations. The effect of repetition and the number of observers were also investigated.

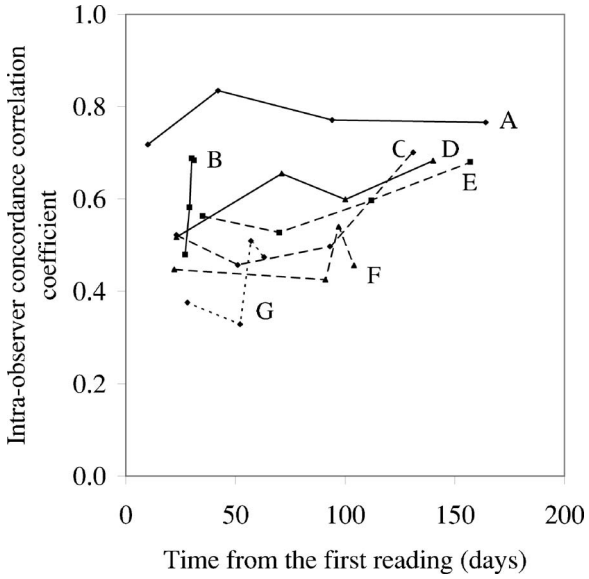

FIG. 2. Change in intraobserver correlation coefficients between two consecutive readings for each of seven observers with five readings.

III. RESULT

When observers were asked to participate in the study for multiple times, the time between two consecutive studies was varied considerably. Figure 2 shows the change in intraobserver correlation coefficients between two consecutive studies for each of seven observers who have completed the study five times. When the time between two studies was very short, such as less than 5 days for three observers (observers B, F, and G), the correlation coefficients were increased, thus indicating that the observers were more consistent. However, the correlation coefficients between the subsequent studies for observers F and G decreased. It is not known whether the differences in these results were due to experimental variation or affected by memory. To minimize the effect of memory, however, it may be desirable to provide a sufficient time between repeated readings. It is apparent in Fig. 2 that the variation in intraobserver correlations for the first two readings is relatively large. It is interesting to note, however, that the correlation coefficients for the last

TABLE II. Average and range of intraobserver correlation coefficients between first and second readings by a single observer.

	Average correlation coefficients
Breast radiologists (9)	0.51 [0.35, 0.66]
Nonradiologists (8)	0.51 [0.38, 0.72]

two readings became comparable and somewhat higher except for two observers F and G. This result may be due to the effect of learning that the observers might have become more familiar in rating similarity for pairs of images and thus became more consistent. The average intra-observer correlation coefficients between the first and second readings and their ranges for groups of breast radiologists and non-radiologists are shown in Table II. The correlation coefficients between the two "single" readings were not very high, indicating that rating the similarity for pairs of images used in this study was difficult and not reproducible at least for the first two readings. The average intraobserver correlation coefficients were comparable for breast and nonradiologists, which indicates that the reproducibility in rating the similarity by the same observer was not related to the experience in reading mammograms.

The similarity ratings from the multiple readings by the same observer were averaged for each observer to reduce the effect of intraobserver variation. As a result, it is expected that the average ratings by each observer become more reliable. Figure 3 shows the decrease in the standard deviation of ratings as the number of readings by these observers increased. The standard deviation of the ratings by seven observers with five readings was first determined for each of 114 pairs, and then the average and standard deviation of the 114 standard deviations are determined and shown. Based on the F test, there was a statistically significant difference ($P < 0.00001$) between the pooled variances of one and two readings. The result in Fig. 3 indicates that although there might not be much benefit by repeating more than two times, the ratings would be more reliable when each observer provided the ratings multiple times than just once. Table III shows the averages and ranges of interobserver correlation coefficients between all possible pairs of observers in each of the three groups of observers by use of the first, second, and

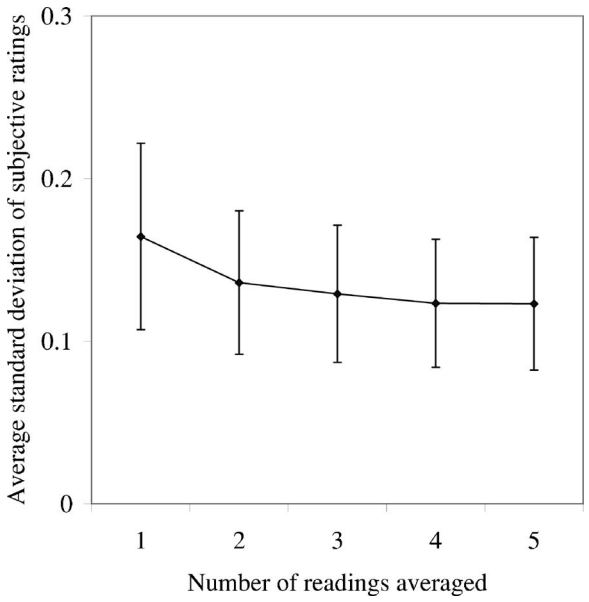

FIG. 3. Effect of the number of readings averaged by the same observer on the average standard deviation of subjective ratings for the observers with five readings.

the average of the two readings. For the groups of breast and nonradiologists, data by the observers with two readings were used. Although interobserver correlations were relatively low, there was a small improvement in the average correlation coefficients for both breast and nonradiologists by taking the average of the two readings. Although the sampling distribution of the correlation coefficients is not normal, the inverse hyperbolic tangent transformation (z transformation) of the correlation coefficients can be assumed to have a normal distribution. If the correlation coefficients were assumed independent random samples, the mean of the correlations for the averaged readings were significantly higher ($P < 0.0001$) than those for the first and second readings based on the paired t test. The correlation coefficients for the group of general radiologists were lower than the other two groups. The reason for this result may be related to the fact that general radiologists had a wide range of experience in reading mammograms, and most of them had no experience in participating observer studies. Table IV shows the average standard deviations of ratings within the group of

TABLE III. Average and range of interobserver correlation coefficients within the group of observers for first and second readings and average of two readings. Correlations were determined for the observers with at least two readings for breast and nonradiologists. P_{12}, P_{1A}, and P_{2A} are P values between the first and second, first and averaged, and second and averaged reading.

	First reading	Second reading	Averaged reading
Breast radiologists (36 combinations)	0.36 [0.16, 0.58]	0.37 [0.14, 0.61] (P_{12}=0.6)	0.47 [0.30, 0.67] ($P_{1A}, P_{2A} < 0.00001$)
General radiologists (45 combinations)	0.25 [0.06, 0.36]	—	—
Nonradiologists (28 combinations)	0.34 [0.10, 0.55]	0.30 [0.07, 0.58] (P_{12}=0.2)	0.41 [0.20, 0.62] ($P_{1A}, P_{2A} < 0.0001$)

TABLE IV. Average standard deviations of ratings within the group of observers for first and second readings and average of two readings. Standard deviations were determined for the observers with at least two readings for breast and non-radiologists.

	First reading	Second reading	Averaged reading
Breast radiologists (9)	0.184	0.179	0.145
General radiologists (10)	0.203	—	—
Non-radiologists (8)	0.180	0.182	0.151

observers. By employing the average of two readings, the interobserver variations were reduced about 16 to 20 %. The results also indicate that the average ratings are more reliable than the single readings. For our purpose of determining reliable similarity ratings by breast radiologists, at least two readings for each observer may be useful.

The similarity ratings for 114 pairs of images by each observer were then averaged for a group of observers. To investigate the effect of the number of observers, a simulation was conducted by use of the first readings by the 13 breast radiologists. Two groups of observers were randomly selected, and the average ratings by the selected observers in each group were determined. Differences in the averaged ratings by two groups were determined for 114 pairs. This process was repeated for 100 times, and the average and standard deviation of the differences were shown in Fig. 4. The average difference in ratings between single observers was as large as 0.23; however, when the number of observers in each group was increased to four, the average difference was reduced about 50% (0.11). The corresponding correlation coefficients between two groups are shown in Fig. 5. The average correlation coefficient was improved from 0.37 with single observers to 0.78 with six observers. The mean of

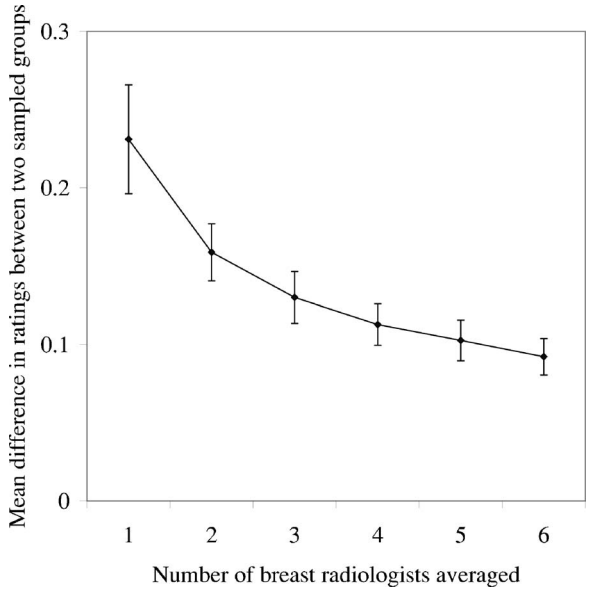

FIG. 4. Effect of the number of observers in each group on the average difference in similarity ratings between two groups.

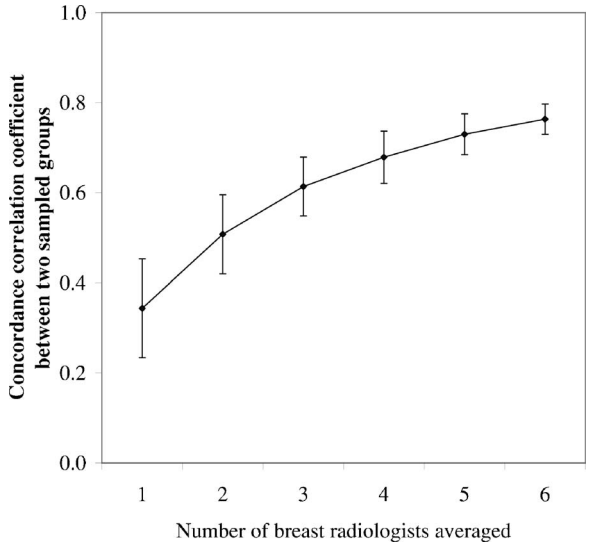

FIG. 5. Effect of the number of observers in each group on the intergroup correlation coefficients.

the correlations with a larger number of observers are all found to be higher ($P<0.00001$) based on two-sample t test. The result suggests that the reliability in similarity ratings would increase as the number of observers increased.

When the subjective similarity ratings were averaged within the groups of breast, general, and nonradiologists, intergroup agreements became very high. Table V shows the correlation coefficients between the groups of the nine breast radiologists and the eight nonradiologists, when first, second, and average ratings were used. The results indicate that the multiple readings by the same observers may be useful in addition to the increase in the number of observers. The relationships between the average ratings by the nine breast radiologists with two readings and ten general radiologists, and by the nine breast and eight nonradiologists with two readings are shown in Figs. 6 and 7, respectively. The correlation coefficient between breast and general, and breast and nonradiologists are 0.846 (95% confidence interval (CI [0.789, 0.888]) and 0.817 (CI [0.747, 0.869]), respectively, which are significantly higher than the correlations between single observers (Table III). The similarity ratings by the general or non-radiologists for some pairs were somewhat different from those of breast radiologists, which were probably due to the difference in diagnostic experience. The pairs of images with a relatively large difference in the av-

TABLE V. Intergroup correlation coefficients between breast and nonradiologists, when first, second, and averaged readings by single observers are averaged within the group. For averaged reading, two readings for each observer are first averaged, and then averaged within the group. The observers with two readings were only included.

	First reading	Second reading	Averaged reading
Nine breast radiologists vs eight nonradiologists	0.736	0.781	0.817

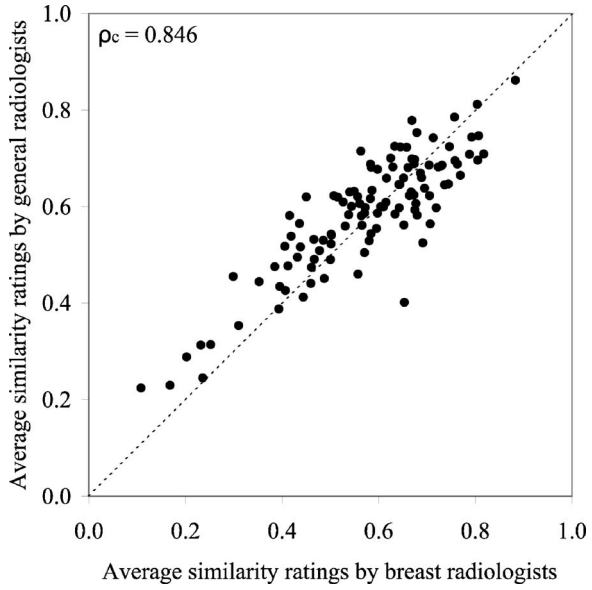

FIG. 6. Relationship between the average ratings by the nine breast radiologists with two readings and ten general radiologists with one reading.

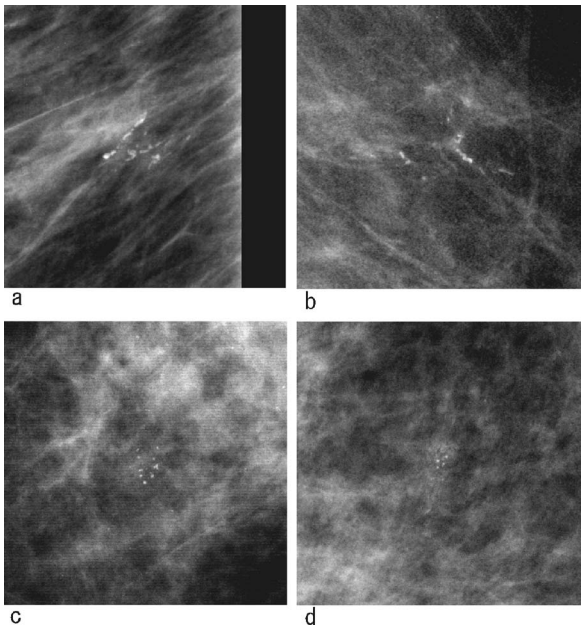

FIG. 8. Pairs of images with relatively large differences in average ratings between breast and general radiologists. Averages and the standard deviations by the nine breast radiologists with two readings and ten general radiologists with one reading are 0.82±0.11 and 0.70±0.15, respectively, for (a) and (b), and 0.65±0.17 and 0.77±0.15, respectively for (c) and (d).

erage ratings between breast and general radiologists are shown in Fig. 8. The top pair [(a) and (b)] was considered very similar by the breast radiologists, whereas the general radiologists found it less similar. On the other hand, the general radiologists considered the second pair [(c) and (d)] more similar, whereas the breast radiologists found it less similar. However, these differences are within or almost within one standard deviation, and therefore, can be considered rather insignificant.

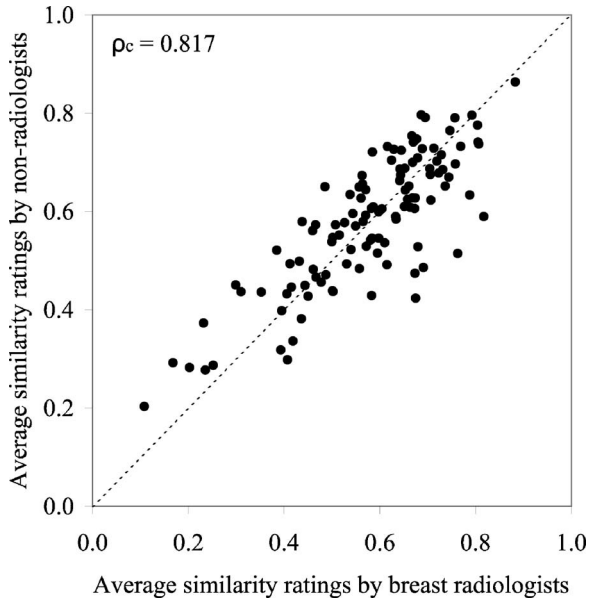

FIG. 7. Relationship between the average ratings by the nine breast and eight nonradiologists with two readings.

IV. DISCUSSIONS

Content-based radiologic image retrieval from Picture Archiving and Communication System (PACS) has been studied by many investigators.[14–21] Depending on the purpose of retrieval, the images retrieved for a query image might be "similar" or the same in terms of pathology, if already known, type of examination, body part, or appearance. For the purpose of diagnostic aid, such as to help radiologists distinguish between benign and malignant lesions, we believe that the retrieved images must be similar in appearance or diagnostic signs from the radiologists' points of view. To our knowledge, limited groups[22–24,31] have examined subjectively whether images are similar based on the similarity ratings provided by radiologists. Li *et al.*[22] have obtained the similarity ratings for pairs of lung nodules in thoracic CT by both radiologists and medical physicists. They found that the average similarity ratings by the group of medical physicists were in good agreement with those by the group of radiologists (Pearson's correlation coefficient 0.88.) The average ratings by the radiologists were considered reliable and were used as a "gold standard" in their study. In our study, radiologists' similarity ratings were obtained for pairs of clustered microcalcifications on mammograms. The variation in observers' impressions of similarity for pairs of clustered microcalcifications could be larger than that of the nodules in thoracic CT, because both characteristics of individual calcifications and the cluster would be considered for diagnosis. For some cases radiologists may strongly consider the distribution of microcalcifications in the cluster, and for others, they may be influenced by some specific features such as the presence of linear microcalcifications. It is possible also that

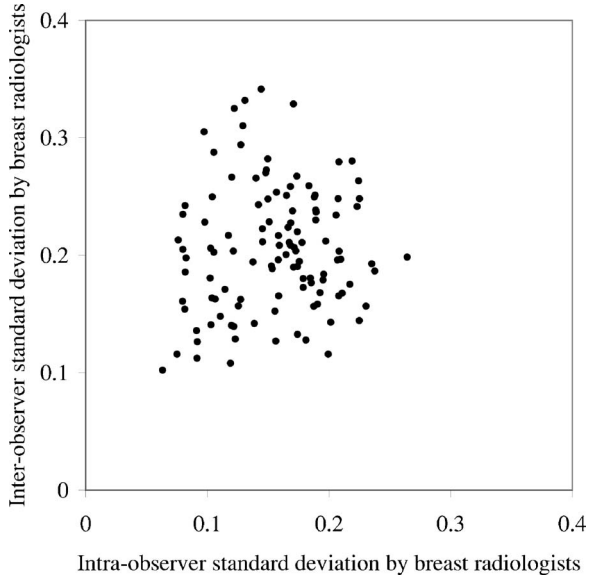

FIG. 9. Relationship between intraobserver and interobserver variations in similarity ratings by nine breast radiologists with two readings.

not all microcalcifications are identified by all observers because of their fine structure. Some observers may find two overlapping microcalcifications, whereas others find it as one microcalcification. Figure 9 shows the relationship between the average intraobserver variation and the interobserver variation in similarity ratings by breast radiologists for 114 pairs. The result shows that for some pairs, the variation between the observers was large although each observer was consistent individually, suggesting that they may be looking at different characteristics. Figure 10 shows a pair of images [(a) and (b)] with both small intra- and small inter-observer variations, and a pair of images [(c) and (d)] with small intraobserver variations but large interobserver variations. The first pair includes relatively small clusters and rather distinct calcifications; however, since the second lesions are relatively large, it is possible that some observers found different signs and features and reacted differently based on their own experiences.

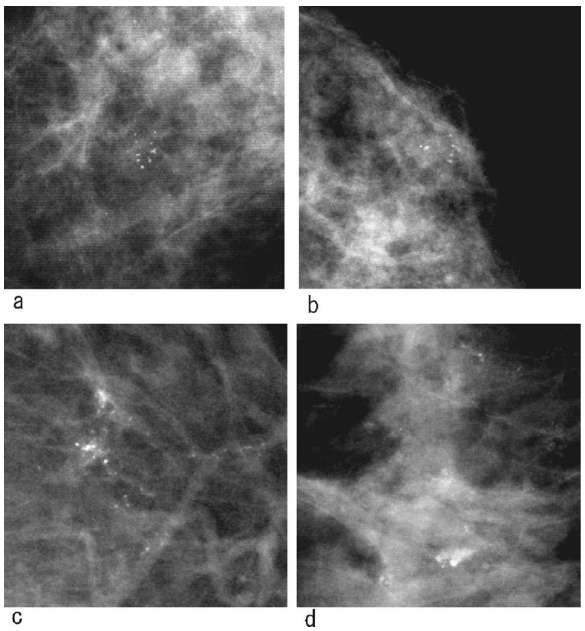

FIG. 10. A pair of images, (a) and (b), with both small intra- and inter-observer standard deviations (0.06 and 0.10, respectively), and a pair of images, (c) and (d), with small intra-observer but large inter-observer standard deviations (0.10 and 0.30, respectively).

Nishikawa *et al.*[23] have investigated two methods, i.e., absolute scale and paired comparison methods, to determine observers' abilities to judge the similarity for pairs of clustered calcifications. In the absolute scale method, 30 pairs of images were shown to the observers one by one, and then the similarity scores from 1 (nearly identical) to 5 (not at all similar) were provided by the observers. In the paired comparison method, the observers compared each pair to all of the other pairs one by one, and marked which pair was more similar than the other. Their result showed that the intraobserver agreement was better with the paired comparison method than the absolute scoring method, whereas the interobserver agreements were comparable. Although observers might be more consistent with the paired comparison method, the similarity score obtained by such a method would be strongly dependent on the cases included in the study. This method is also time consuming because each pair must be compared to a large number of pairs, e.g., 29 pairs in their study (a total of 435 comparisons). In general, a good correlation between the scores for the two methods was found, indicating that similarity of clustered calcifications can be determined in a meaningful way. However, the numbers of cases (30 pairs) and observers (four observers including three breast radiologists) used in their study were rather small.

El-Naqa *et al.*[24] also obtained similarity ratings in an absolute scale for pairs of clustered microcalcifications. In their study, similarity ratings were provided by observers with background in medical image analysis. The criterion for similarity was limited only to the geometric distribution of microcalcifications in the cluster, and the observers read the images with circles overlaid at locations of individual microcalcifications that were previously identified by experts. The observer agreement might be good by limiting the criteria and with the location of each microcalcification marked; however, without the original images, the markings could be very eye distracting and all of the other important features for diagnosis of lesions were not considered. We believe that similar images retrieved by use of such similarity ratings may not be helpful for distinction of benign and malignant lesions.

In this study, we asked observers to rate the similarity based on the overall impression for diagnosis. For similar images to be useful to radiologists in their diagnosis, images to be presented to radiologists must be similar in terms of an overall diagnostic point of view. Therefore, we believe that this criterion was important for the determination of subjective similarity ratings. We have employed an absolute scale method with a continuous rating scale from 0 to 1 to obtain the subjective similarity ratings. Presentation of six pairs si-

multaneously may help observers to scale their impression because they can compare six "known" images and decide which "known" images are more similar or less similar to the "unknown" image. On the other hand, while the similarity ratings would not be completely independent, they would not be too strongly dependent on cases included as in paired comparison or ranking methods. The numbers of cases used in this study was rather small; a larger number of cases would be needed to include various types of lesions for the development of CAD schemes. In this study, high correlation coefficients between the average subjective ratings by two groups of observers were obtained. Although there are variations in subjective impression within and between individuals, the statistical variation can be reduced by obtaining the data from a number of observers and their repeated readings. The high correlation between the groups of observers indicated that a component of impression of similarity for pairs of images may be commonly shared by human observers, and reliable similarity ratings can be obtained by this method. We believe that average similarity ratings by experienced radiologists determined in this way are meaningful and useful for determination and evaluation of objective similarity measures in CAD schemes.

ACKNOWLEDGMENTS

This work was supported by USPHS Grant No. CA61625. The authors are grateful to H. Abe, M.D., Ph.D., F. Li, M.D., Ph.D., H. Nishide, M.S., and H. Arimura, Ph.D. for valuable discussion, and to the following for their participation in the observer study: C. Sennett, M.D., J. Chambliss, M.D., M. Linver, M.D., G. Cardenosa, M.D., G. W. Eklund, M.D., E. Mendelson, M.D., J. A. Wolfman, M.D., L. I. Segal, M.D., T. Kuritza, D.O., Y. T. Adler, M.D., T. Stroud, M.D., A. S. Fiore, M.D., A. Favelukes, M.D., E. Groskind, M.D., J. Molin, M.D., L. R. Dale, M.D., P. Dadros, M.D., T. M. Goodnight, M.D., J. A. Zuckerman, M.D., A. Okamura, M.D., A. Edward, M.A., J. Papaioannou, M.S., M. Yaffe, Ph.D., and E. Hendrick, Ph.D. K. D. and R. A. S. are shareholders of R2 Technology, Inc., Los Altos, CA. CAD technologies developed in the Kurt Rossmann Laboratories have been licensed to companies including R2 Technology, Riverain Medical, Mitsubishi Space Software Co., General Electric Corporation, Medican Technologies and Toshiba Corporation. It is the policy of the University of Chicago that investigators disclose publicly actual or potential significant financial interests that may appear to be affected by research activities.

APPENDIX

Unknown	Known	Unknown	Known	Unknown	Known	Unknown	Known	Unknown	Known
0087LM3	0171RC1	0126RM1	0057RC1	0411LC1	0309RC1	0473RM1	0325LM1	0476LM1	0012RM1
	0325LC1		0511LC1		1465RM1		1503RC1		0276RC1
	1124LM1		1213LM1		1807RC1		1850RM1		1213LC1
	1245LC1		3030RC1		1839LC1		3486RC1		1376LM1
	1743LC1		3367LC1		1916LM1		4098LM1		3459RC1
	1924RC1		4159LM1		3026RC4		4179RC1		3502RC1
0488RM1	1227LC1	1115RC1	1232RM1	1175RM1	0285RM1	1176LM1	0285RC1	1261LC1	0503LM1
	1465RM1		1332LC1		1465RC1		0315RM1		1438LM1
	1837LC1		1605LM1		1601RM1		0335RC1		1729RC1
	1916LM1		1809LM1		1619RC1		1175RC1		1743LC1
	3121RM1		3502RC1		1774LC1		1223LM1		4161RM1
	4162LC1		4105RC1		4147LC1		1619RM1		4179RC1
1448LM1	0012RM1	1530LC1	1452RM1	1840LC1	1382RM1	1866LC1	0057RC1	1934LM1	0236RC1
	1214LC1		1601RM1		1647LC1		0167RM1		0276RC1
	1431RM1		1619RC1		1721LC2		1175RC1		1191LC1
	1913LC1		1850RC1		1729RC1		1406LM1		1223LM1
	3044LC1		1850RM1		1766LC1		3367LC1		1530LM1
	3499LC1		4099LM1		4171LM1		4196RM1		3367LC1
3037LC1	0151RM1	3361LM1	0344LM1	3436RC1	0400LC2	4151RM1	0288LM1		
	0511LM1		1201RM1		1797LM1		1213LC1		
	1632LM1		1482RM1		1874RM1		1406LC1		
	1743LC1		1913LM1		3007LM1		3507LM2		
	1894RM1		1944RC1		1009RM1		4179RC1		
	3037LM1		3516LM1		3400LM1		4196RM1		

Note: The first four digits represent the case number in the DDSM followed by the breast (R: right or L: left), view (C: CC or M: MLO), and the lesion number.

[1]American Cancer Society, *Cancer Facts and Figures 2006* (American Cancer Society, Atlanta, 2006).
[2]S. V. Destounis, P. DiNitto, W. Logan-Young, E. Bonaccio, M. L. Zuley, and K. M. Willlison, "Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience," Radiology **232**, 578–584 (2004).
[3]D. Gur, J. S. Stalder, L. A. Hardesty, B. Zheng, J. H. Sumkin, D. M. Chough, B. E. Shindel, and H. E. Rockette, "Computer-aided detection performance in mammographic examination of masses: Assessment," Radiology **233**, 418–423 (2004).
[4]R. F. Brem, J. Baum, M. Lechner, S. Kaplan, S. Souders, L. G. Naul, and J. Hoffmeister, "Improvement in sensitivity of screening mammography with computer-aided detection: A multiinstitutional trial," AJR, Am. J. Roentgenol. **181**, 687–693 (2003).
[5]H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement inradiologists' detection of clustered microcalcifications on mammograms," Invest. Radiol. **25**, 1102–1110 (1990).
[6]R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," Radiology **219**, 192–202 (2001).
[7]S. A. Butler, R. J. Gabbay, D. A. Kass, D. E. Siedler, K. F. O'Shaughnessy, and R. A. Castellino, "Computer-aided detection in diagnostic mammography: detection of clinically unsuspected cancers," AJR, Am. J. Roentgenol. **183**, 1511–1515 (2004).
[8]T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center," Radiology **220**, 781–786 (2001).
[9]Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," Acad. Radiol. **5**, 155–168 (1998).
[10]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and L. Hadjiiski, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. **25**, 516–526 (1998).
[11]I. Leichter, S. Buchbinder, P. Bamberger, B. Novak, S. Fields, and R. Lederman, "Quantitative characterization of mass lesion on digitized mammograms for computer-assisted diagnosis," Invest. Radiol. **35**, 366–372 (2000).
[12]H. P. Chan, B. Sahiner, D. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic microcalcifications morphological and texture feature space," Med. Phys. **25**, 2007–2019 (1998).
[13]Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: Automated feature analysis and classification," Radiology **198**, 671–678 (1996).
[14]H. A. Swett, P. R. Fisher, A. I. Cohn, P. L. Miller, and P. G. Mutalik, "Expert system-controlled image display," Radiology **172**, 487–493 (1989).
[15]H. A. Swett, P. G. Mutalik, V. P. Neklesa, L. Horvath, C. Lee, J. Richter, I. Tocino, and P. Fisher, "Voice-activated retrieval of mammography reference images," J. Digit Imaging **11**, 65–73 (1998).
[16]G. Bucci, S. Cagnoni, and R. De Dominicis, "Integrating content-based retrieval in a medical image reference database," Comput. Med. Imaging Graph. **20**, 231–241 (1996).
[17]S. T. C. Wong and H. K. Huang, "Design methods and architectural issues of integrated medical image data base systems," Comput. Med. Imaging Graph., **20** 285–299 (1996).
[18]U. Sinha and H. Kangarloo, "Principal component analysis for content-based image retrieval," Radiographics **22**, 1271–1289 (2002).
[19]H. Qi and W. E. Snyder, "Content-based image retrieval in picture archiving and communications systems," J. Digit Imaging **12**(2), 81–83 (1999).
[20]A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C. R. Shyu, and A. Marchiori, "Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment," Radiology **228**, 265–270 (2003).
[21]M. L. Giger, Z. Huo, C. J. Vyborny, L. Lan, I. Bonta, K. Horsch, R. M. Nishikawa, and I. Rosenbourgh, "Intelligent CAD workstation for breast imaging using similarity to known lesions and multiple visual prompt aids," Proc. SPIE **4684**, 768–773 (2002).
[22]Q. Li, F. Li, J. Shiraishi, S. Katsuragawa, S. Sone, and K. Doi, "Investigation of new psychophysical measures for evaluation of similar images on thoracic CT for distinction between benign and malignant nodules," Med. Phys. **30**, 2584–2593 (2003).
[23]R. M. Nishikawa, Y. Yang, D. Huo, M. Wernick, C. A. Sennett, J. Papaioannou, and L. Wei, "Observers' ability to judge the similarity of clustered calcifications on mammograms," Proc. SPIE **5372**, 192–198 (2004).
[24]I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, "A similarity learning approach to content-based image retrieval: Application to digital mammography," IEEE Trans. Med. Imaging **23**, 1233–1244 (2004).
[25]M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer, Jr., "Current status of the Digital Database for Screening Mammography," *Digital Mammography* (Kluwer Academic, Dordrecht, 1998), pp. 457–460.
[26]L. I. K. Lin, "A concordance correlation coefficient to evaluate reproducibility," Biometrics **45**, 255–268 (1989).
[27]L. I. K. Lin, "A note on the concordance correlation coefficient," Biometrics **56**, 324–325 (2000).
[28]J. M. Bland and D. G. Altman, "Measuring agreement in method comparison studies," Stat. Methods Med. Res. **8**, 135–160 (1999).
[29]P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," Psychol. Bull. **86**, 420–428 (1979).
[30]C. A. E. Nickerson, "A note on "A concordance correlation coefficient to evaluate reproducibility," Biometrics **53**, 1503–1507 (1997).
[31]C. Muramatsu, Q. Li, K. Suzuki, R. A. Schmidt, J. Shiraishi, G. M. Newstead, and K. Doi, "Investigation of psychophysical measure for evaluation of similar images for mammographic masses: preliminary results," Med. Phys. **32**, 2295–2304 (2005).