Determination of subjective similarity for pairs of masses and pairs of clustered microcalcifications on mammograms: Comparison of similarity ranking scores and absolute similarity ratings

Chisako Muramatsu,[a] Qiang Li, Robert A. Schmidt, Junji Shiraishi, Kenji Suzuki, Gillian M. Newstead, and Kunio Doi
Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, The University of Chicago, 5841 South Maryland Avenue, MC 2026, Chicago, Illinois 60637

(Received 8 February 2007; revised 17 April 2007; accepted for publication 10 May 2007; published 19 June 2007)

The presentation of images that are similar to that of an unknown lesion seen on a mammogram may be helpful for radiologists to correctly diagnose that lesion. For similar images to be useful, they must be quite similar from the radiologists' point of view. We have been trying to quantify the radiologists' impression of similarity for pairs of lesions and to establish a "gold standard" for development and evaluation of a computerized scheme for selecting such similar images. However, it is considered difficult to reliably and accurately determine similarity ratings, because they are subjective. In this study, we compared the subjective similarities obtained by two different methods, an absolute rating method and a 2-alternative forced-choice (2AFC) method, to demonstrate that reliable similarity ratings can be determined by the responses of a group of radiologists. The absolute similarity ratings were previously obtained for pairs of masses and pairs of microcalcifications from five and nine radiologists, respectively. In this study, similarity ranking scores for eight pairs of masses and eight pairs of microcalcifications were determined by use of the 2AFC method. In the first session, the eight pairs of masses and eight pairs of microcalcifications were grouped and compared separately for determining the similarity ranking scores. In the second session, another similarity ranking score was determined by use of mixed pairs, i.e., by comparison of the similarity of a mass pair with that of a calcification pair. Four pairs of masses and four pairs of microcalcifications were grouped together to create two sets of eight pairs. The average absolute similarity ratings and the average similarity ranking scores showed very good correlations in the first study (Pearson's correlation coefficients: 0.94 and 0.98 for masses and microcalcifications, respectively). Moreover, in the second study, the correlations between the absolute ratings and the ranking scores were also very high (0.92 and 0.96), which implies that the observers were able to compare the similarity of a mass pair with that of a calcification pair consistently. These results provide evidence that the concept of similarity for pairs of images is robust, even across different lesion types, and that radiologists are able to reliably determine subjective similarity for pairs of breast lesions. © *2007 American Association of Physicists in Medicine*. [DOI: 10.1118/1.2745937]

Key words: mammograms, computer-aided diagnosis, similar images, observer study

I. INTRODUCTION

Breast cancer is the most frequently diagnosed non-skin cancer and the second leading cause of cancer deaths in women in the United States. According to the American Cancer Society,[1] 240 510 new cases were expected in 2007, including DCIS. Mammography is currently considered the most useful screening method for early detection of breast cancers in the general population. Earlier studies of randomized clinical trials[2–4] have shown that periodic mammographic screening can reduce the breast cancer mortality. However, for detecting cancers at an early, favorable stage, many patients with benign lesions are also sent for biopsy. In fact, only about 15%–40% of the lesions that are biopsied based on the mammographic findings are found to be malignant.[5–7] To improve this low positive biopsy yield, investigators have been developing computerized schemes in the hope of aiding radiologists in the distinction between benign and malignant lesions on mammograms. The results of receiver operating characteristic studies have indicated that computer-aided diagnosis (CAD), in which radiologists make diagnoses by taking into consideration the computer output, has the potential to improve the diagnosis of masses[8,9] and clustered microcalcifications.[10] In these studies, observers were provided with the likelihood of malignancy of the lesions as a numerical percentage. Some radiologists may use this quantitative likelihood effectively; others may not do so because of the lack of understanding as to why a computer estimates lesions to be suspicious or likely to be benign.

Another way to help radiologists is to provide images with known pathology, which are similar to that of an unknown lesion. Presentation of similar images as a computer aid has been studied by several investigators for interpretation in chest radiography,[11] thoracic computed tomography (CT),[12,13] and mammography.[14–19] Some of these studies[12,13,16,17,20] demonstrated that the presentation of simi-

lar images has the potential to improve radiologists' diagnostic performance. In order to retrieve images that are helpful, we believe that the images must be really similar to an unknown lesion, as judged from the radiologists' points of view. Consequently, it is important to obtain similarity ratings for many pairs of lesions by many radiologists. However, it is considered difficult to reliably determine the subjective similarity of images, and some variations among radiologists are expected.

In this study, image similarities determined by two different methods are compared; (1) an absolute rating method and (2) a 2-alternative forced-choice (2AFC) method, also known as a paired comparison method. Our goal was to demonstrate that reliable and useful similarity ratings for pairs of masses and pairs of clustered microcalcifications can be determined. A similar study was conducted by Nishikawa *et al.*,[21] in which they investigated the similarity for 30 pairs of calcifications determined by the two methods, i.e., the absolute method and the paired comparison method. Four observers, including three breast radiologists and one experienced research technician, participated in their study. In general, there was a good correlation between the scores for the two methods (Pearson's correlation coefficient of −0.77); however, there were substantial differences for some pairs. One of the disadvantages of the 2AFC method is that the results obtained are only rankings, and they may not have a linear relationship with the absolute ratings. For example, pairs with ranks of one and two and pairs with ranks of two and three are both considered as the same difference in one rank, even if the difference in similarities of the former pairs may be larger than that of the latter pairs. In this study, we selected only eight pairs each for masses and microcalcifications. The purpose was to demonstrate simply and clearly the relationship between the similarities obtained by the two methods. The number of observers was increased to ten to reduce the effect of inter-observer variability and the effect of the quantized ranking. In addition, another set of similarity ranking scores was determined by use of mixed pairs, i.e., by comparison of the similarity of a mass pair with that of a calcification pair. To our knowledge, no one has investigated whether the similarities of pairs of different types of lesions can be compared. Our hypothesis was that if observers do employ a basic concept of similarity for pairs of lesions, the similarity of a mass pair could be compared with that of a calcification pair consistently.

II. MATERIAL AND METHODS

The images used in this study were obtained from the publicly available database, Digital Database for Screening Mammography,[22] developed by the University of South Florida. Regions of interest (ROIs) 5 cm by 5 cm in size were obtained for benign and malignant masses and microcalcifications, all of which had been verified by biopsy. The contrast and density level for each ROI were manually adjusted to appropriate level by a breast radiologist to facilitate the visual comparison.

A. Previous studies with an absolute rating method

In our previous study on masses,[23] subjective similarity ratings were obtained for 60 pairs of masses by radiologists. Our purpose in the previous study was to determine the subjective ratings for pairs of masses and to use them as a "gold standard" for the development of an objective similarity measure for selection of similar images in our CAD scheme. First, ten ROIs with five benign and five malignant masses were selected as "unknown" images. Those images were selected to include masses with different sizes and various characteristics. For each unknown image, six images were selected as "known" images to be paired with the unknown image so that the expected similarity ratings would range from very dissimilar to very similar by subjective judgment. Five observers, including two breast and three general radiologists participated in the study.

We have also conducted another observer study to obtain the subjective similarity ratings for 114 pairs of microcalcifications.[24] Nineteen ROIs with ten benign and nine malignant microcalcifications were first selected as unknown images that included different sizes and various types of microcalcifications. For each unknown image, six known images were selected in a similar way. In the calcification study, a total of 33 observers including 13 breast radiologists, ten general radiologists, and ten non-radiologists participated, and some observers repeated the study multiple times.

During these studies, an unknown image was placed in the center of a monitor, and then six known images, each paired with the unknown image, were placed on the right and left. The six known images as well as the unknown sets were randomized, and the observers were blinded to their pathologies. The observers were asked to provide the six similarity ratings on a continuous rating scale between 0 and 1, corresponding to "not similar at all" and "almost identical," respectively, based on the overall impression for diagnosis. We considered these to be absolute ratings, because the ratings of 0 and 1 were given specific meanings, and the rating for each pair could be determined subjectively without other pairs.

These two studies formed the nucleus for our current study. For the mass pairs, the average ratings by the five radiologists were considered as the gold standard, and for the calcification pairs, the average ratings by the nine breast radiologists who provided the ratings twice were considered as the gold standard.

B. Determination of similarity ranking score by use of a 2AFC method

In the current 2AFC study, eight pairs of masses and eight pairs of microcalcifications from the previous studies were used. The eight pairs were selected so that their similarity ratings from the previous studies were approximately evenly distributed (a difference of about 0.1 between two pairs), and the standard errors of the ratings by the five and nine observers for masses and microcalcifications, respectively, were approximately equal to or less than 0.05. In this way, the eight pairs could be distinguished or "ranked" based on the absolute ratings. No image was used more than once.

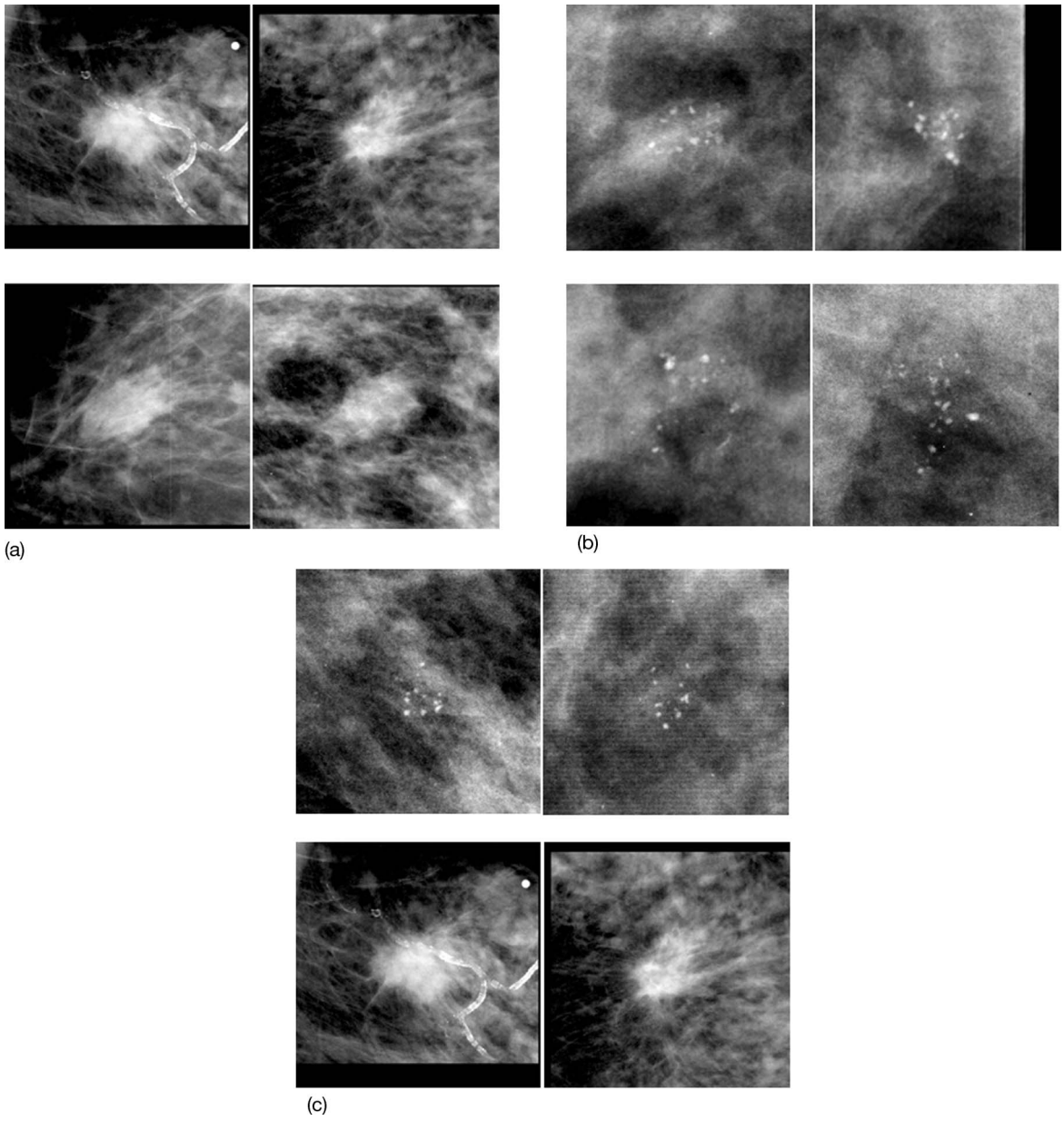

FIG. 1. Cases for the first 2AFC study with (a) the mass pairs, which were ranked second and third and (b) the calcification pairs, which were ranked second and third, and (c) the case for the second 2AFC study with the first-ranked calcification pair and the second-ranked mass pair.

During the study, each pair was compared with all of the other seven pairs one by one; therefore, the same two pairs appeared twice during the study. The observers were asked to compare the similarities of two pairs and to select one pair that they considered more similar than the other based on the overall impression for diagnosis. The number of times that a pair was selected as the more similar one was counted and considered as the similarity ranking score for the pair. For each pair, two scores were determined by use of the first and second rounds for investigating the intra-observer consistency. The averages of the two scores were used for the other analysis. The order and the positions (above or below) were randomized, and the positions were switched when the same two pairs appeared for the second time, to reduce the effect of the location.[25] Ten observers, including four breast radiologists, one breast imaging fellow, two general radiologists, and three radiology residents, participated in the 2AFC study. The four breast radiologists and one general radiologist also participated in the previous mass and/or calcification studies. The average ranking scores by the ten observers were compared with the absolute similarity ratings from the previous studies.

There were two sessions in the 2AFC study. In the first 2AFC study, the eight mass pairs and eight calcification pairs were grouped separately. Figures 1(a) and 1(b) show the comparisons of the mass pairs, which were ranked second (upper pair) and third (lower pair) based on the absolute ratings, and of the calcification pairs, which were ranked second (upper) and third (lower), respectively. In the second 2AFC study, the four mass pairs with the odd numbers of

similarity ranks based on the absolute ratings were combined with the four calcification pairs with the even numbers of ranks, and vice versa, to create two sets of eight pairs. Figure 1(c) shows the case when the first-ranked calcification pair (upper) was compared with the second-ranked mass pair (lower). The second study was conducted for investigating whether the concept of similarity can be realized even if different types of lesion pairs are compared.

III. RESULTS

In the first 2AFC study, there was a total of 56 comparisons (28 comparisons each for the mass set and the calcification set). On average, the observers were consistent in choosing the same pair when the pair was compared with another pair twice. The average number of times that they selected the same pairs were 25.1 (90%) and 25.2 (90%) for the mass and calcification sets, respectively. The average Spearman's rank ordered correlation coefficients between the similarity ranking scores for the first and second readings by the same observers (intra-observer correlations) were 0.92 (range, [0.84,0.98]) and 0.90 [0.63,0.98] for the mass and calcification sets, respectively. Although the first and the second readings were obtained within one session, many observers commented that they remembered seeing the same case, but did not remember which pair they chose the first time. When the similarity ranking scores by the same observers were averaged, the average Pearson's correlation coefficients between all possible pairs of observers (inter-observer correlation) were 0.74 [0.14,0.96] and 0.86 [0.54,0.99] for the mass and calcification sets, respectively. For the mass set, one observer selected more similar pairs differently; however, most of the other observers agreed well with each other. For the calcification set, another observer selected more similar pairs a little differently; however, the other nine observers agreed with each other very well. Although these two observers were not experienced breast radiologists, there did not seem to be an obvious trend in the different levels of experience. The average similarity ranking scores for the 16 pairs of lesions were determined by averaging of the scores of the ten observers. Figures 2(a) and 2(b) show the relationship between the absolute similarity ratings and the similarity ranking scores by the 2AFC method for the mass and calcification sets, respectively. Although one mass pair was ranked higher by the 2AFC method than by the absolute rating method, and the ranks of the fourth and fifth similar calcification pairs were reversed, there were very good correlations between the similarities by the two methods. The correlation coefficients between the average absolute ratings and the average ranking scores were 0.94 and 0.98 for the mass and calcification pairs, respectively.

In the second 2AFC study, the mass pairs and calcification pairs were mixed to create two sets of eight pairs, whose absolute similarity ratings were not so evenly distributed, but ranged from very dissimilar to very similar. The average numbers of times that the observers selected the same pair the first and second time were 25.8 (92%) and 25.5 (91%) for the two sets. The average intra-observer rank-ordered correlation coefficients were 0.95 [0.88,0.98] and 0.95 [0.88,1.00] for the two sets. However, the agreement between the observers was not as good as that in the first study. The average inter-observer correlation coefficients between all the possible pairs of observers were 0.77 and 0.51 for the two sets. There were three observers, including the two "outliers" in the first study, who selected more similar pairs differently for the second set. However, when the similarity ranking scores of the ten observers were averaged, the average ranking scores agreed well with the absolute similarity ratings. The relationship between the average absolute similarity ratings and the average similarity ranking scores for the two sets are shown in Figs. 3(a) and 3(b). As in the first study, the same mass pair was ranked higher by the 2AFC method than by the absolute rating method. The Pearson's correlation coefficients between the similarities by the two methods were 0.92 and 0.96.

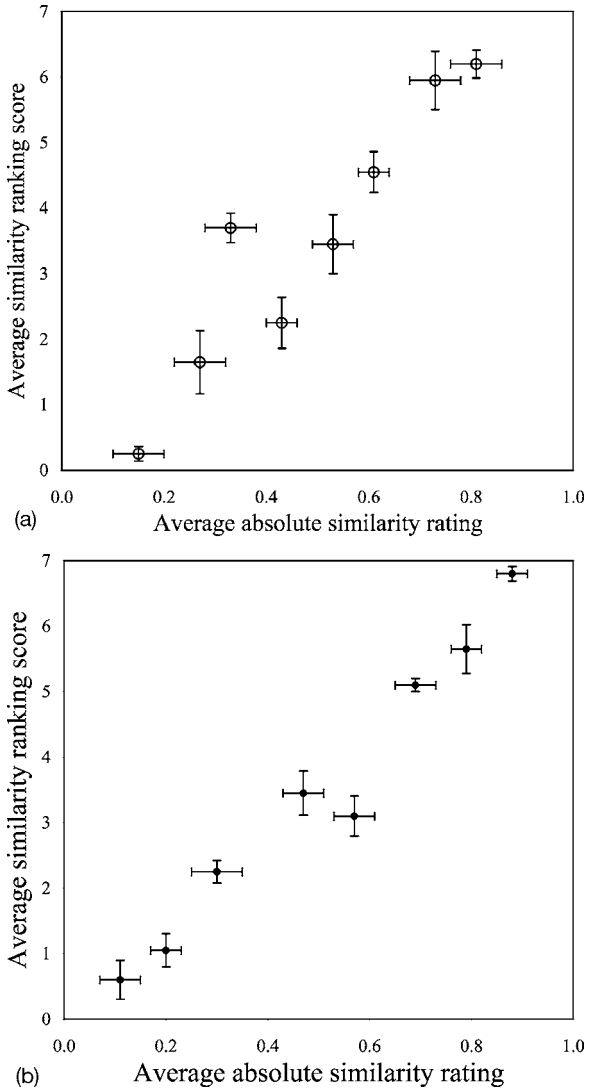

FIG. 2. Relationships between the average absolute similarity ratings obtained previously and the average similarity ranking scores determined by the 2AFC method for (a) the mass pairs and (b) the calcification pairs.

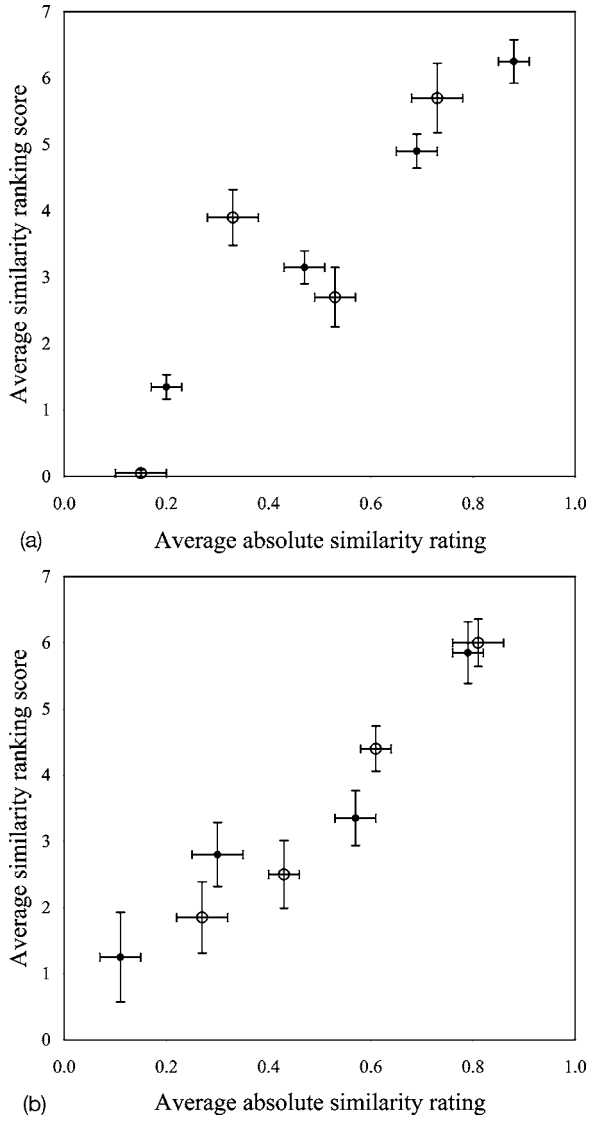

FIG. 3. Relationships between the average absolute similarity ratings obtained previously and the average similarity ranking scores determined by the second 2AFC study for two sets, (a) and (b), when the mass pairs were compared with the calcification pairs.

IV. DISCUSSION

The presentation of images with known pathology similar to that of a new unknown lesion would be useful if radiologists considered them as similar. In order to develop a computerized scheme, which would find useful images from a large database, it is important to obtain the data on radiologists' impression of similarity for a large number of pairs of lesions. In our previous studies,[23,24] subjective similarity ratings for pairs of masses and for pairs of microcalcifications were obtained from radiologists in an absolute scale. Overall, we believe that we could obtain reliable data, as indicated by the high correlations between the average ratings by the different groups of observers. However, it can be difficult to determine the absolute similarity for pairs of breast lesions. In fact, the average inter-observer correlation between pairs of breast radiologists were not very high for the calcification pairs (Pearson's correlation coefficient, $r=0.51$ with the averages of two readings). On the other hand, it is expected that human observers are generally more consistent in choosing one or the other by comparison of the two, e.g., choosing a more similar pair than the other. By use of the 2AFC method, a pair is compared to all other pairs one by one, and the number of times the pair was chosen over the other can be counted and considered as a measure of the similarity score. However, one of the disadvantages of the 2AFC method is that this method only provides rankings, and therefore, the scores would strongly depend on the cases used. Another disadvantage is that using the 2AFC method takes time if the number of cases is large. For the purpose of our determining a gold standard for development of a computerized scheme, therefore, it is desirable to obtain similarity ratings in an absolute scale. In our previous studies, six pairs of images containing the same unknown image were presented to the observers simultaneously, so that the observers could compare and "scale" their ratings. This presentation method, we believe, could help observers to be more consistent in rating within the six pair set; however, it was not known whether the observers were consistent among the different sets.

To investigate whether similarity ratings for pairs of masses and for pairs of microcalcifications can be determined reliably, we conducted a simple experiment in which we determined the similarity ranking scores by the 2AFC method with 16 selected pairs. Although we selected only eight pairs each, the similarity difference of 0.1 is rather subtle, as shown in Fig. 1, and we were not sure that the observers could judge the differences. However, the results showed good agreement between the average similarity ranking scores by the 2AFC method and the average absolute similarity ratings from the previous studies.

After the first study, we questioned whether a similarity rating of 0.7, for example, for a mass pair would have a comparable meaning for a calcification pair. If so, it would be possible to compare the absolute similarity ratings for different types of lesions as well as the similarity ranking scores that are determined by use of mixed pairs. However, it was not known whether observers could compare the similarities of a mass pair to that of a calcification pair in a consistent manner. In our previous studies,[23,24] the observers seemed to have more difficulty in rating the calcification pairs than rating the mass pairs. In the reading of microcalcifications, radiologists consider not only the cluster distribution, but also the shape of individual microcalcifications. If the concept of similarity is robust, a consistent result would be expected even by use of mixed pairs. Therefore, we conducted a second study. When the mass pairs and calcification pairs were mixed in the second study, the differences in the absolute ratings of some of the two adjacent-ranked pairs were small; therefore, this was considered more challenging. However, the result was very encouraging. The good correlations between the similarity ranking scores and the absolute ratings indicate that the observers share the basic concept of similarity for pairs of breast lesions, and that the subjective similarity can be determined in the absolute scale in a consistent manner. We believe that the similarity ratings

can be determined reliably and can be used as the gold standard for the development of a computerized scheme.

ACKNOWLEDGMENTS

This work was supported by USPHS Grant No. CA62625. The authors are grateful to F. Li, M.D., Ph.D., and S. Kasai, Ph.D., for valuable discussions, and to the following for their participation in the observer study: C. Sennett, M.D., H. Abe, M.D., Ph.D., D. Berger, M.D., F. Li, M.D., Ph.D., A. Shimauchi, M.D., Ph.D., S. Zangan, M.D., A. Bennett, M.D., and M. Cranford, M.D. K. Doi and R. A. Schmidt 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 Company, General Electric Corporation, Median 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.

[a]Reprint requests to: Chisako Muramatsu, Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, The University of Chicago, 5841 S. Maryland Ave MC 2026, Chicago, IL 60637. Phone: (773) 834-5094, Fax: (773) 702-0371. Electronic mail: chisa@uchicago.edu

[1]American Cancer Society. *Cancer Facts and Figures 2007* (American Cancer Society, Atlanta, 2007).

[2]L. Tabar, G. Fagerberg, S. W. Duffy, N. E. Day, A. Gad, and O. Grontoft, "Update of the Swedish two-county program of mammographic screening for breast cancer," Radiol. Clin. North Am. **30**, 187–210 (1992).

[3]S. Shapiro, W. Venet, P. Strax, L. Venet, and R. Roeser, "Selection, follow-up, and analysis in the health insurance plan study: A randomized trial with breast cancer screening," JNCI, J. Natl. Cancer Inst. **67**, 65–74 (1985).

[4]L. L. Humphrey, M. Helfand, B. K. S. Chan, and S. H. Woolf, "Breast cancer screening: a summary of the evidence for the U.S. preventive services task force," Ann. Intern Med. **137**, E-347–E-367 (2002).

[5]F. M. Hall, J. M. Storella, D. Z. Silverstone, and G. Wyshak, "Nonpalpable breast lesions: Recommendations for biopsy based on suspicion of carcinoma at mammography," Radiology **167**, 353–358 (1988).

[6]D. B. Kopans, R. H. Moore, K. A. McCarthy, D. A. Hall, C. A. Hulka, G. J. Whitman, P. J. Slanetz, and E. F. Halpern, "Positive predictive value of breast biopsy performed as a result of mammography: There is no abrupt change at age 50 years," Radiology **200**, 357–360 (1996).

[7]E. A. Sickles, D. L. Miglioretti, R. Ballard-Barbash, B. M. Geller, J. W. T. Leung, R. D. Rosenberg, R. Smith-Bingman, and B. C. Yankaskas, "Performance benchmarks for diagnostic mammography," Radiology **235**, 775–790 (2005).

[8]Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: Effectiveness of computer-aided diagnosis-observer study with independent database of mammograms," Radiology **224**, 560–568 (2002).

[9]H. P. Chan, B. Sahiner, M. A. Helvie, N. Patrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An ROC study," Radiology **212**, 817–827 (1999).

[10]Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," Acad. Radiol. **6**, 22–33 (1999).

[11]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).

[12]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).

[13]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).

[14]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).

[15]H. Qi and W. E. Snyder, "Content-based image retrieval in picture archiving and communications systems," J. Digit Imaging **12** (2 Suppl. 1), 81–83 (1999).

[16]J. Sklansky, E. Y. Tao, M. Bazargan, C. J. Ornes, R. C. Murchison, and S. Teklehaimanot, "Computer-aided, case-based diagnosis of mammographic regions of interest containing microcalcifications," Acad. Radiol. **7**, 395–405 (2000).

[17]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).

[18]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).

[19]B. Zheng, A. Lu, L. A. Hardesty, J. H. Sumkin, C. M. Hakim, M. A. Ganott, and D. Gur, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," Med. Phys. **33**, 111–117 (2006).

[20]K. Horsch, M. L. Giger, C. J. Vyborny, L. Lan, E. B. Mendelson, and R. E. Hendrick, "Classification of breast lesions with multimodality computer-aided diagnosis: Observer study results on an independent clinical data set," Radiology **240**, 357–368 (2006).

[21]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).

[22]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 Publishers, Dordrecht, 1998), pp. 457–460.

[23]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).

[24]C. Muramatsu, Q. Li, R. A. Schmidt, K. Suzuki, J. Shiraishi, G. M. Newstead, and K. Doi, "Experimental determination of subjective similarity for pairs of clustered microcalcifications on mammograms: Observer study results," Med. Phys. **33**, 3460–3468 (2006).

[25]R. L. Day, "Position bias in paired product tests," J. Mark. Res. **6**, 98–100 (1969).