Investigation of psychophysical measure for evaluation of similar images for mammographic masses: Preliminary results

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

(Received 21 December 2004; revised 5 April 2005; accepted for publication 11 May 2005; published 20 June 2005)

We investigated a psychophysical similarity measure for selection of images similar to those of unknown masses on mammograms, which may assist radiologists in the distinction between benign and malignant masses. Sixty pairs of masses were selected from 1445 mass images prepared for this study, which were obtained from the Digital Database for Screening Mammography by the University of South Florida. Five radiologists provided subjective similarity ratings for these 60 pairs of masses based on the overall impression for diagnosis. Radiologists' subjective ratings were marked on a continuous rating scale and quantified between 0 and 1, which correspond to pairs not similar at all and pairs almost identical, respectively. By use of the subjective ratings as "gold standard," similarity measures based on the Euclidean distance between pairs in feature space and the psychophysical measure were determined. For determination of the psychophysical similarity measure, an artificial neural network (ANN) was employed to learn the relationship between radiologists' average subjective similarity ratings and computer-extracted image features. To evaluate the usefulness of the similarity measures, the agreement with the radiologists' subjective similarity ratings was assessed in terms of correlation coefficients between the average subjective ratings and the similarity measures. A commonly used similarity measure based on the Euclidean distance was moderately correlated ($r=0.644$) with the radiologists' average subjective ratings, whereas the psychophysical measure by use of the ANN was highly correlated ($r=0.798$). The preliminary result indicates that a psychophysical similarity measure would be useful in the selection of images similar to those of unknown masses on mammograms. [DOI: 10.1118/1.1944913]

Key words: mammograms, breast masses, computer-aided diagnosis, similar image, artificial neural network, observer study

I. INTRODUCTION

Breast cancer is the second leading cause of cancer deaths for women in the United States. According to the American Cancer Society, the number of new breast cancer cases in the United States in 2004 is estimated to be 275 380, with 215 990 of these invasive.[1] Although mammography is considered at present to be the most useful screening method for early detection of breast cancer, it is difficult to detect abnormalities on mammograms, especially for patients with dense breasts. Moreover, when an abnormality is detected, it can be very difficult even for experienced radiologists to determine whether the lesion is malignant or benign. In fact, the positive predictive value of biopsies recommended based on mammograms is usually in the low range, from 15% to 35%.[2–5]

To assist radiologists in reducing the number of biopsies performed on benign lesions, investigators have attempted to develop computer-aided diagnostic schemes for classification of breast lesions. With these schemes, the likelihood of malignancy of breast lesions would be determined, and the computer results would be presented to radiologists as a second opinion. Huo *et al.*[6] developed a computerized scheme[7–9] for classification of mammographic masses that involves extraction of features related to the margin and the density of the masses. The image features were determined not only from standard views, but also from a special view such as a spot-compression view. Their study showed that most radiologists were able to improve their classification accuracy with the computer aid. The area (Az) under the average receiver operating characteristic (ROC) curve was improved from 0.93 to 0.96 with the computer aid, although the Az value of the computerized scheme alone (0.90) was lower than those for most radiologists without the computer aid.

With computer-aided diagnosis (CAD), radiologists read images and make a diagnosis by taking into consideration the "second opinions" provided by computers. When radiologists find that the computer output is in agreement with their diagnoses, they may become more confident in their decisions. On the other hand, when radiologists disagree with the computer output and have confidence in their decisions, they would probably ignore the computer output. However, if radiologists are not very confident and consider that the computer output might be correct, they might reconsider their decisions. In this way, we expect that radiologists could uti-

lize the computer output beneficially, and that the result could be superior to the diagnosis either by radiologists or by the computer alone.

In order for radiologists to utilize CAD effectively, the output from computers must be easily interpretable. With CAD for detection of breast lesions, radiologists are prompted to reexamine suspicious areas so that they can decide whether or not lesions are present. On the other hand, with CAD for classification of breast lesions, the likelihood of malignancy would be represented by numerical values, which most radiologists do not encounter in clinical practice. The numerical values alone may not be enough to convince radiologists about the likelihood of malignancy, because the reasons for lesions being highly suspicious for malignancy (or benignity), such as a spiculated (or sharp) margin, or an irregular (or round) shape, are not provided. Therefore, unlike computer aids for *detection* of lesions, it can be difficult to utilize computer aid for the *classification* of lesions.

Chan *et al.*[10] conducted an observer study to compare radiologists' performance in the distinction between benign and malignant mass lesions without and with the aid of their computerized scheme.[11] With the computer aid, the performance of only two out of six radiologists was equal to or greater than that of the computer (Az value of 0.92) for classification of 238 masses. Jiang *et al.*[12] investigated the usefulness of their CAD scheme[13] in assisting radiologists in the classification of clustered microcalcifications on mammograms. Although the average Az value for ten radiologists improved from 0.61 to 0.75 with the computer aid, it was still lower than the Az value (0.80) for the computer analysis.

Radiologists are commonly trained to distinguish between breast cancers and benign lesions on mammograms by reading many malignant and benign cases. Therefore, one may assume that a radiologist has a large database in his or her brain which may consist of the cases that he or she has encountered in textbooks and in clinical practice. If the radiologist is faced with a new unknown lesion in daily practice, he or she may attempt to recall some previous cases that are similar to the unknown lesion, and then these previous cases should be helpful to the radiologist in making a clinical decision. Therefore, we believe that the presentation of a set of malignant and benign images similar to those of the unknown lesion would be very helpful to radiologists in addition to numerical values such as the likelihood of malignancy.

For development of such a computerized scheme, a measure of similarity must be defined so that one can find similar images from a large database. The similar images selected by the computerized scheme need to be really similar from a radiologist's point of view. If a radiologist has the impression that the images selected are not similar, these images will not be helpful for diagnosis. However, it is difficult to quantify radiologists' impressions on similarity for a pair of images and to determine the objective similarity measure that would agree with radiologists' subjective similarity ratings. If radiologists could provide experimental data for subjective similarity ratings for a variety of pairs of masses, then it might be possible to determine a similarity measure that would agree well with radiologists' visual impression.

Li *et al.*[14] investigated a new method for selection of similar nodules on thoracic CT, which could be used as an aid for distinction between malignant and benign nodules. They developed a new similarity measure, called a psychophysical measure, for lung nodules by use of an artificial neural network (ANN) which was employed to learn the relationship between radiologists' subjective similarity ratings and the image features for pairs of nodules. They found that the psychophysical similarity measure correlated well with radiologists' average subjective ratings (correlation coefficient $r=0.72$). However, it is uncertain whether the concept of a psychophysical measure can be employed for mammographic mass lesions, because the characteristics of lesions and those of normal structures on mammograms are quite different from those in thoracic CT. In this study, we investigated a psychophysical similarity measure for mass lesions on mammograms that may be useful for selection of similar images.

II. MATERIAL AND METHODS

A. Database for mass lesions in mammograms

The mass images used in this study were obtained from the Digital Database for Screening Mammography (DDSM),[15] which is organized by the University of South Florida and publicly available via website.[16] The DDSM consists of 914 cancer, 996 benign, and 695 normal cases, each of which nominally includes four digitized standard-view mammograms obtained from the examinations conducted at four facilities from 1988 to 1999.[15] For our study, 5 cm by 5 cm regions of interest (ROIs) were extracted for biopsy-proven mass lesions, if the outlines of the lesions provided in the DDSM did not exceed the size of the ROI. ROIs were obtained from both cranio-caudal and mediolateral oblique views if both were specified in the DDSM; some lesions (4%) were apparent and/or marked only in one view. In this study, we employed only ROIs that included a mass with no other lesion visible within the same ROI; if a mass was partially visible at the edge of an image, the case was not included. Because this study was focused on mass lesions, we did not include lesions which were considered to be architectural distortion or asymmetric density findings, nor those suspected to be lymph nodes by a breast radiologist (R.A.S.). Other images that were not suitable for this study, such as unclear lesions, lesions with markings, and lesions with a skin fold, were also excluded. For our preliminary study, only the cases digitized to a 12 bit grey scale were retrieved; other cases will be included in a future study.

With the above exclusion criteria, about 24% of cases in the DDSM were removed, and the images used in this study consisted of 681 malignant and 764 benign ROIs. According to the Breast Imaging Reporting and Data System (BI-RADS) descriptions provided in the DDSM, the shapes of the mass lesions employed in our study were round (9%), oval (24%), lobulated (27%), and irregular (32%), and the

degrees of subtlety for the majority (84%) of the lesions were between 3 and 5, where 5 corresponds to least subtle. The effective diameters of the masses ranged from 4 to 37 mm, with a mean of 14 mm. The DDSM contained images that were digitized with three different pixel sizes of 42, 43.5, and 50 μm. The matrix size for all of the ROIs was therefore adjusted to 500 by 500 pixels, with a pixel size of 100 μm, by linear interpolation and downsampling. For facilitating visual comparisons by radiologists, the contrast and density levels for all of the ROIs were adjusted subjectively to appropriate levels under the guidance of a breast radiologist (R.A.S.). These adjusted images were also used for determination of image features.

B. Observer study for determination of subjective similarity ratings

An observer study was conducted to obtain radiologists' subjective similarity ratings for pairs of mass lesions. Five radiologists, including two experienced breast radiologists (20 and 12 years of experience) and three general radiologists (18, 18, and 15 years of experience) participated in this study. First, ten (five malignant and five benign) masses were selected as "unknown" cases by a stratified randomization scheme from three groups of small, medium-sized, or large lesions. The effective diameters of the selected masses ranged from 8 to 20 mm. For each of the ten unknown cases, three malignant and three benign masses were selected as similar or dissimilar cases, thus providing a total of 60 pairs of masses. These similar or dissimilar cases must be distributed in a wide range of similarity. If the cases were selected randomly, most of pairs would be dissimilar, and would not be useful in this study. First, several candidates for similar or dissimilar images were selected by use of simple image features such as size and contrast. The selected candidates of similar or dissimilar images for each unknown image were then categorized subjectively into similar, somewhat similar, or not similar image by one of the co-authors (C.M.). For each unknown case, the final selection of cases was made with inclusion of two images in each of the three categories. The cases were carefully selected so that they were not strongly dependent on some specific features. No more than one ROI from the same patient was employed.

During the observer study, six pairs of mass lesions were shown on a liquid crystal display (LCD) monitor (CCL314, 20.8 in., 2048×1536 pixels, 200 cd/m^2 white luminance; Totoku Electric Co., Ltd., Tokyo, Japan) where the unknown case was placed in the center with three similar or dissimilar cases each on the right and left side. The order of unknown cases as well as of similar or dissimilar cases was randomized, and the pathologies of the lesions were not revealed to observers. Zooming and windowing capabilities were provided, although most observers did not change the window setting as it had already been subjectively optimized. The instructions to observers included: (1) The purpose of this study was to obtain basic data for selecting similar images in a CAD scheme to assist radiologists' interpretation of mammograms; (2) cases included ten unknown masses (five malignant and five benign), each with six similar or dissimilar masses (three malignant and three benign); (3) subjective similarity should be marked based on the overall impression for radiologic diagnosis on a continuous rating scale[17] between 0 and 1 with a line-checking method, where 0 and 1 correspond to "two masses not similar at all" and "two masses almost identical," respectively; (4) the rating should be given independently and consistently; and (5) the reading time was not limited. The average ratings from the five radiologists were employed for data analysis.

Before this observer study, a preliminary test was conducted for assessment of the feasibility of an observer study in which observers, including 7 radiologists and 16 nonradiologists (9 medical physicists, 5 medical physics students, and 2 technical staff members), were asked to provide the level of subjective similarity for pairs of mass lesions that were printed on films. Subjective ratings were recorded on a 6.3 cm continuous rating scale and quantified from 0.00 to 1.00 as described earlier by use of a ruler. Six pairs of masses (5 cm $\times$ 5 cm ROIs) were printed on one film in the above described format. A total of 60 pairs of mass lesions, which included 43 pairs of masses used in the subsequent study, were evaluated. At that point, images that were later considered inappropriate and subsequently were excluded from this study as explained in the previous section were still included, and the contrast and density level were suboptimal. Therefore, the results from the two studies cannot be compared. The result of this preliminary test showed, however, that the average subjective ratings by radiologists agreed well with that by nonradiologists (correlation coefficient 0.864). The high correlation indicated that, although radiologists have special skills in reading radiologic images, the basic concept of the similarity of two images may be commonly shared by human observers, and the subjective impression for similarity can be quantified reliably with this method. Because all observers in this study participated in the preliminary test with the films approximately five to seven months earlier, a training session was not provided in this study with the LCD monitor.

C. Determination of image features for mass lesions

For accurate determination of features, the outlines of mass lesions were traced manually by one of two radiologists who were not aware of the pathologic diagnoses of the cases. To examine the variation in outlines provided by different radiologists, three radiologists including the above two radiologists were asked to draw outlines for a limited number of cases. Although there were some variations for some of the cases, we considered that the difference was rather insignificant, and the effect on determination of feature was not very large. Six morphological features were determined on the basis of outlines, and seven grey-level features and thirteen edge-gradient features were determined in four regions, i.e., a region inside the outline, a strip-like adjacent region inside the outline, a strip-like adjacent region outside the outline, and two adjacent regions combined, as shown in Fig. 1. The morphological features included the effective diameter,

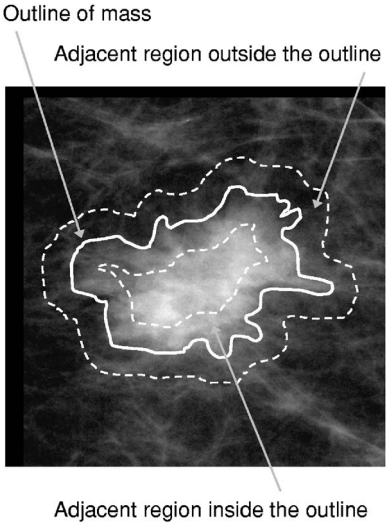


FIG. 1. ROI for a mass and regions in which features are determined.

which is defined by the diameter of a circle that has the same area as that of the mass lesion; the degree of circularity,[18,19] which is defined by the fraction of the overlapped area of the circle with the mass; and the degree of irregularity,[18,19] which is defined by 1 minus the ratio of the perimeter of the circle to the length of the lesion outline. Other morphological features were the minor-to-major-axis ratio (MMR) of an ellipse fitted to the lesion outline, and the degree of ellipticity and the elliptical irregularity,[19] which were defined in the same manner as the degrees of circularity and irregularity, respectively, by use of a fitted ellipse instead of the circle.

The grey-level features included the mass contrast, the mean contrast, the edge contrast, the standard deviations (SDs) in pixel values, and the fractions of overlapped area of normalized grey-level histograms in two regions.[19] The mass contrast was defined by the difference between the mean pixel value of 7×7 pixels around the centroid of the lesion and the pixel value at 2% from the lower end of the histogram inside the outline. Two percent was employed empirically so that the effect of noise was limited. The mean contrast was defined by the difference in the mean pixel values between the region inside the outline and the adjacent region outside the outline, whereas the edge contrast was defined by the difference in the mean pixel values between the two adjacent regions inside and outside the outline. Three features were determined based on the SD, i.e., the SD in terms of the absolute pixel value, the SD normalized by the mean pixel value, and the SD normalized by the mean contrast. The fractions of overlapped area were determined in the region inside the outline and in the adjacent region outside the outline, and also in the adjacent regions inside and outside the outline.

The features based on edge characteristics included the radial gradient indexes (RGI),[7] the modified radial gradient index, which is based on the fitted ellipse, the mean magnitude of the gradient, and features that characterize cumulative maximum edge-gradient histograms[7] for both radial and modified radial angles. The histogram-related features were the full width at half maximum (FWHM), the fraction of the area under the histograms between 135° and 225°, the SD of radial angles, the width at half area under the histogram centered at 180°, and the ratio of the mean to the peak value of the histogram. The magnitude and direction of edge-gradients were determined by use of a 5×5 Sobel-like filter.[20]

D. Determination and evaluation of similarity measures by use of the differences in feature values and the Euclidean distance

In order to determine an objective similarity measure between two masses, their feature values were compared. When the feature values were very close, i.e., when the difference in a feature value was very small, two masses would be considered similar according to that feature. However, the similarity rating by radiologists for the pair might or might not be high. If the differences in feature values for two masses were highly correlated with radiologists' similarity ratings, the feature would be considered effective in determining an objective similarity measure. In order to facilitate the comparison between subjective ratings and objective measures, the difference in feature values was converted to the range from 0 to 1 by use of an exponential function,[14] so that 1 corresponds to the feature values of two masses being the same, and 0 corresponds to the feature values being extremely different. The correlation coefficients between the subjective ratings and objective measures were determined for the pairs of masses employed in this observer study, so that the usefulness of individual features as well as their combinations for determination of similarity measures could be examined. For combinations of selected features, the Euclidean distance in the multiple-dimensional feature space between two masses was employed.

E. Determination and evaluation of psychophysical similarity measure

Our psychophysical similarity measure was based on the idea of relating physical measures, i.e., image features, to radiologists' subjective similarity ratings, which are based on their knowledge and experience in the diagnosis of cancer. The psychophysical measure was obtained by training of an ANN[21,22] with the subjective similarity ratings as teacher data and the pairs of the feature values for two masses as input data.[14] Three-layered, feed-forward network with a backpropagation algorithm was used. Once the ANN was trained by use of training cases, and when the set of corresponding feature values for a new pair of masses was entered to the trained ANN, the psychophysical similarity measure would be provided as the output, expected to be in good agreement with the radiologists' subjective similarity rating.

Results from the ANN were evaluated by use of a round-robin test method. In the round-robin method for this study, six pairs of masses in which one of the pairs corresponding to one unknown image were excluded, and the remaining 54 pairs were used in the training. It is known that a large number of cases are usually required in the training of an ANN in

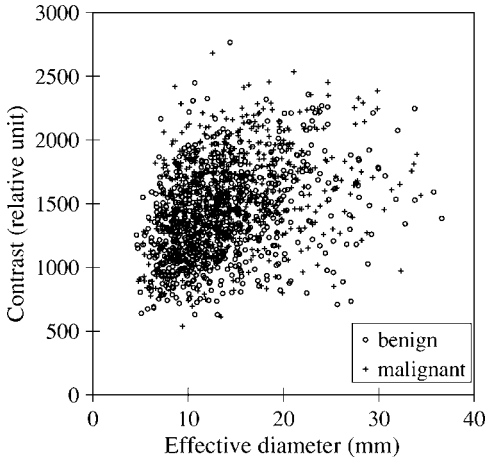


FIG. 2. Relationship between the effective diameter and the contrast of the masses used in this study.

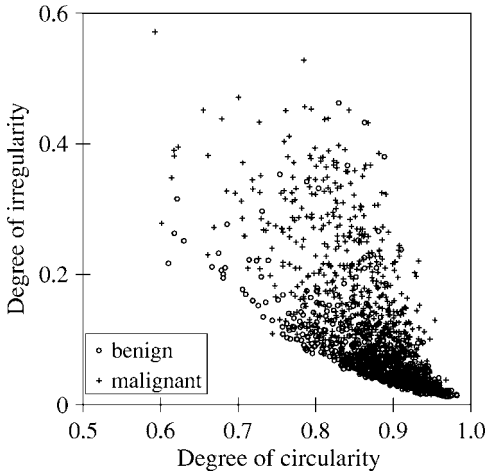


FIG. 3. Relationship between the degree of circularity and the degree of irregularity for the masses used in this study.

order for the trained ANN to be generalized to "unknown" test cases. Moreover, when the ANN is applied to solving a regression problem, such as the one in this study, instead of a classification problem, the trained ANN may not exhibit an optimal performance for test cases, the output of which may be above the highest or below the lowest of output values in the training cases. Therefore, to scale the ANN output, teacher data for both ends, i.e., a set of identical pairs and another set of extremely dissimilar pairs corresponding to the very high and very low subjective ratings, respectively, were created and included in the training set. The numbers of these teacher data was made small (1/3 to 1/6) compared to the number of actual pairs. The trained ANN then provided the psychophysical measure for the six pairs excluded from the training. This process was repeated for all 10 sets of 6 pairs one by one. The usefulness of the psychophysical similarity measures was also evaluated by the correlation coefficients between subjective ratings and the psychophysical measures, in addition to the mean error of the ANN outputs compared with the teacher data.[21] With the mean error, not only the linear relationship between the two, but also how close the results were from the subjective ratings was examined to find whether the ANN was trained adequately.

III. RESULTS

Figure 2 shows the relationship between the effective diameter and the contrast of 1445 masses used in this study. The result shows that the effective diameter and the contrast of the masses are distributed widely. In addition, because the image contrast and the density level were adjusted subjectively, there is no simple relationship between the effective diameter and the contrast. In these features, the distribution for benign masses is very similar to that for malignant masses. Figure 3 shows the relationship between the degree of circularity and the degree of irregularity of the masses. Although malignant masses are likely to have a lower degree of circularity and a higher degree of irregularity than benign masses, most malignant and benign masses cannot be separated by these features alone. These results indicate that, for a given unknown case, there is a good chance to find both malignant and benign similar masses in this large database, because most masses can be found close to other masses in terms of these feature values.

Figure 4 shows the subjective ratings for 60 pairs of masses by two breast radiologists, obtained in the observer study. The diagonal line indicates the perfect agreement. The result shows an interobserver variation in subjective similarity ratings between the two breast radiologists. The correlation coefficient for these two radiologists was high ($r=0.745$). However, there were some variations in subjective ratings among five radiologists, and the correlation coefficients between any two radiologists (there were 10 pairs of radiologists) ranged from 0.415 to 0.745 (mean of 0.527). Although the correlation coefficients between two observers might be relatively small, the reliability of the ratings would be increased by employing average ratings from several observers. In fact, the correlation coefficients between the average ratings of two of five radiologists and the average rat-

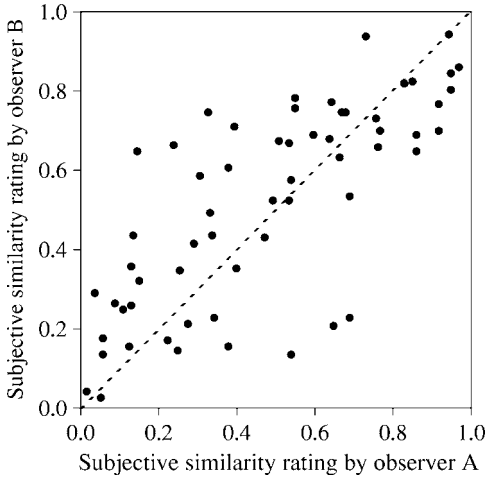
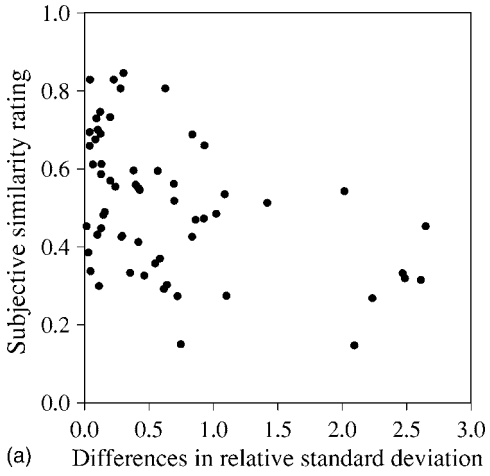


FIG. 4. Result from the observer study showing interobserver variability of two breast radiologists in subjective similarity ratings for 60 pairs of masses (correlation coefficient $r=0.745$)



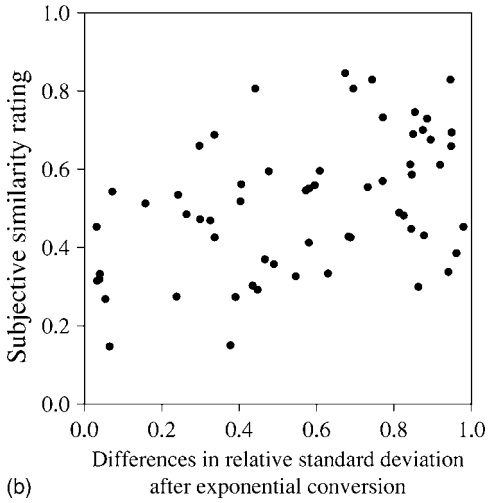


FIG. 5. (a) Relationship between the subjective ratings and differences in relative standard deviation in pixel values for 60 pairs of masses (before conversion). (b) Relationship between the subjective ratings and differences in relative standard deviation in pixel values for 60 pairs of masses (after exponential conversion).

TABLE I. Correlation coefficients between subjective similarity ratings and the differences in one feature within pairs for 60 pairs of masses.

Feature	Correlation coefficient
Standard deviation in pixel values normalized by the mean contrast in adjacent region outside the outline	0.444
Standard deviation in pixel values in adjacent region outside the outline	0.430
Standard deviation in pixel values normalized by the mean pixel value in region inside the outline	0.384
Mass contrast	0.326
Elliptical irregularity	0.278
Effective diameter	0.263
Modified FWHM in two adjacent regions	0.249
Circularity	0.213
Mean contrast	0.193
Irregularity	0.162
Minor-to-major axis ratio of fitted ellipse	0.101
Radial gradient index in adjacent region outside the outline	0.086

ings of the remaining three radiologists were improved, and ranged from 0.616 to 0.829 (mean of 0.737). Therefore, for this study, the average ratings by five radiologists were employed as the "gold standard" for subsequent data analysis.

When only one feature was employed for determination of an objective similarity measure, the highest correlation coefficient between the average subjective ratings and the objective measure was obtained by use of the standard deviation of pixel values normalized by the mean contrast. Figures 5(a) and 5(b) show the relationships between the radiologists' average subjective similarity ratings and the objective measure based on this feature before and after the use of the exponential conversion, respectively. The result indicates a weak correlation between the subjective ratings and the objective measure, and the data points are widely spread. Table I shows the correlation coefficients between the subjective ratings and objective measures for the features which either provided higher correlation coefficients when only one feature was employed or were useful when combined with other features. Although the correlation coefficients were quite low when only one feature was employed, these features may be useful for determination of the objective measures when the other features were combined.

When multiple features were employed for determination of an objective similarity measure, the correlation coefficient between the radiologists' average subjective ratings and the objective measure became as high as 0.644 [95% confidence interval(CI), (0.466, 0.772)], as shown in Fig. 6. The dotted line in Fig. 6 indicates the ideal relationship that we would like to accomplish. Although we determined the correlation coefficient as a measure to evaluate the agreement, we also considered that the data points be distributed adjacent to the dotted line, i.e., close to the "gold standard." The objective measure was determined based on six features, namely, circularity, MMR, mean contrast, SD in the adjacent region outside the outline, SD normalized by the mean pixel value in the region inside the outline, and the modified FWHM in two adjacent regions combined. By use of this objective

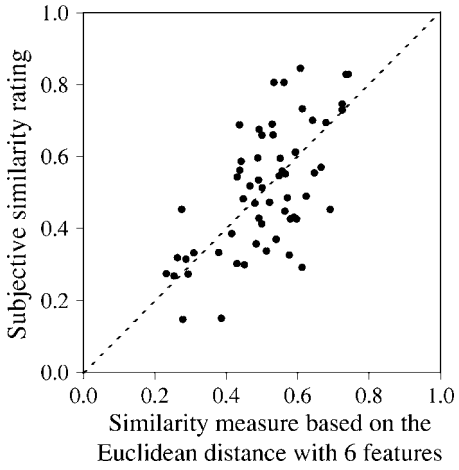


FIG. 6. Relationship between radiologists' average subjective ratings and objective similarity measures by use of the Euclidean distance based on six features.

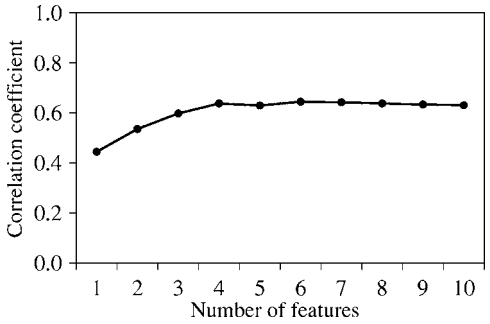


FIG. 7. Correlation coefficients between radiologists' average subjective ratings and objective measures by the combination of features with the highest correlation for different numbers of features used.

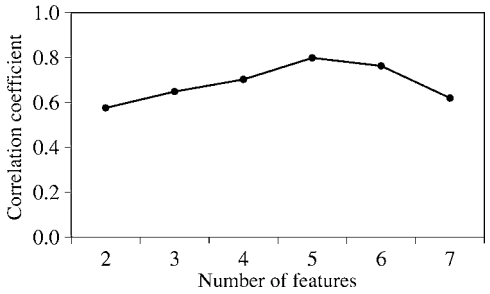


FIG. 8. Correlation coefficients between radiologists' average subjective ratings and psychophysical measures with the best results for different numbers of features used.

measure, some of the selected similar images might or might not be similar to an unknown image, and might or might not be useful to radiologists for diagnosis of an unknown image. Figure 7 illustrates the relationship between the number of features used for determination of objective measures and the correlation coefficients between the subjective ratings and the objective measures. For each number of features, we considered all possible combinations of features up to six features and adding all possible features one by one to the previous combination thereafter. With the different combinations of features, the correlation coefficients varied considerably, and the highest value obtained for each number of features is shown in Fig. 7. As the number of features was increased, the correlation coefficient between the subjective ratings and objective measure was increased at the beginning, gradually saturated, and then decreased. This is probably due to some correlations between some of the features, so that combining more features did not provide any additional useful information.

Psychophysical similarity measures were determined by training of the ANN with different numbers of features and different combinations of various features. The combinations of features which provided relatively high correlations with the average subjective ratings by use of the Euclidean distance were tested. We also tested the combinations of features, each of which provided a relatively higher correlation with the subjective ratings. For each combination of features employed for training of the ANN, the result was tested by use of a round-robin method, and the correlation coefficient between the subjective ratings and the psychophysical measure was determined. Figure 8 shows the correlation coefficient of the best result when the correlation between the subjective ratings and psychophysical measures and the mean error were considered for each number of features employed. As the number of features was increased, the correlation coefficient first increased and then started to decrease. The result was similar to that for the objective measures by use of the Euclidean distance. As a result, a combination of five features provided the best result, which included MMR, SD in the adjacent region outside the outline, SD normalized by the mean pixel value inside the outline, modified FWHM in two adjacent regions combined, and RGI in the adjacent region outside the outline. Figure 9 shows the relationship between the subjective ratings and the psychophysical measure. The psychophysical similarity measure was highly correlated [$r=0.798$, CI (0.682, 0.875)] with the subjective similarity ratings. For the scaling purposes, identical pairs and dissimilar pairs were employed in the training of the ANN. The identical pairs were created from the 63 masses that were included in the 54 pairs of training cases, and dissimilar pairs were created from the independent cases in the database by selecting the pairs that were very far from each other in the Euclidean feature space. There were some variations in ANN training when the identical pairs were created randomly from the independent cases in the database or when dissimilar pairs were created from the training cases. However, in each situation, the best results in terms of the correlation coefficient were comparable, in the range from 0.779 to 0.805.

IV. DISCUSSION

Investigators have suggested the presentation of similar images to help radiologists in the diagnosis of disease by studying different methods for selection of similar cases. Swett *et al.*[23,24] have developed computer-based expert systems, called IMAGE/ICON and MAMMO/ICON, to help in the diagnosis of lung diseases and breast cancer. The system retrieves similar images based on findings in the textual re-

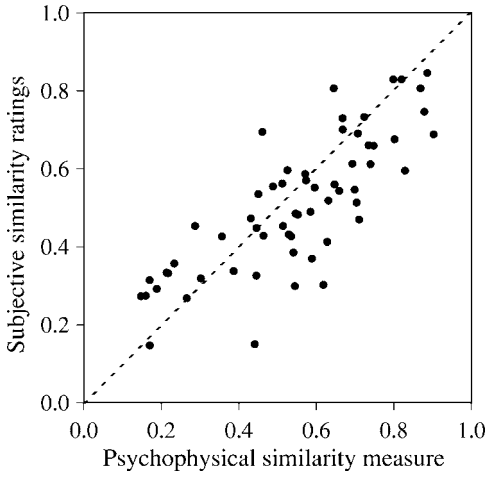


FIG. 9. Relationship between radiologists' average subjective ratings and psychophysical measure by use of five features.

port or in the dictation, which is translated by a speech recognition system. Qi *et al.*[25] proposed content-based image retrieval, in which similar images would be selected by their contents, such as texture, color, or shape. In application to mass lesions on mammograms, the shape of a mass was characterized by a feature vector, and the images with a small vector distance from a query image were retrieved. Aisen *et al.*[26] have reported the potential usefulness of a storage and retrieval system by use of a pattern-recognition technique for thin-section thoracic computed tomography. When the system is presented with an unknown query image, the system predicts the disease for the unknown image, and then known images in the predicted disease category are displayed. Giger *et al.*[27] developed an intelligent CAD workstation for breast lesions. In their system, similar images on mammograms and sonograms are searched based on a single feature, multiple features, or the computer estimate of the likelihood of malignancy. However, we believe that all of these methods have a limitation, because the textual descriptions and computer-extracted features or patterns do not take into account radiologists' visual impression of "similarity" of two images. Therefore, it is questionable whether these retrieved images would be really similar visually. Our study is unique in that radiologists' impression of similarity for a pair of lesions is quantified, and then used as teaching data for determination of a similarity measure. We expect that images selected as similar to an unknown image by our method would be more similar visually than those selected by other methods. Recently, a study similar to ours has been presented by Nishikawa *et al.*[28] in which they compared two methods, namely, the use of absolute scale and paired comparison, for determination of subjective similarity for pairs of clustered microcalcifications. In general, they found good agreement between the two methods. In our study, an absolute scale was also employed with six pairs shown at the same time. By presenting image pairs this way, observers could compare six images and scale their impression. Although the similarity ratings would not be completely independent, we believe that the ratings would not be too strongly dependent on the cases included as with the paired comparison method.

Figure 10 shows three pairs of masses in which the psychophysical measure agreed well with the radiologists' average subjective similarity ratings. The three pairs of masses represent very similar (0.83), somewhat similar (0.57), and not very similar (0.15) masses. When an objective similarity measure was determined by use of the Euclidean distance, the similarity measures for the very similar pair, the somewhat similar pair, and the not very similar pair were 0.61, 0.66, and 0.28, respectively. Therefore, by use of the Euclidean distance, these three pairs would be considered somewhat similar (first two pairs) or not similar (third pair), and such an objective measure might not be useful for the three pairs in Fig. 10. The difference in these results between the psychophysical measure and the Euclidean distance indicates that not only the differences in feature values, but also the feature values themselves may be important for determination of a reliable similarity measure, and the ANN could be trained effectively by use of the pairs of feature values. Figure 11 shows two pairs of masses in which the difference between the subjective similarity ratings and psychophysical measure was relatively large. The top pair was considered similar (0.69) by radiologists; however, the psychophysical measure (0.46) was not very high. On the other hand, the psychophysical similarity measure (0.90) was very high for the bottom pair, whereas radiologists' average rating (0.69) was somewhat lower. There are two possible reasons for this discrepancy. First, radiologists' impression may not be completely reflected in the features which were used for training the ANN. For the top pair, radiologists may think that both masses have a relatively smooth and clear margin. However, because of the overlapped tissue, this may not be accurately accounted for in determination of the feature values. Additional features or better methods for determination of features may be needed. Second, subjective ratings were obtained for only 60 pairs of masses, and these 60 pairs may not have included a number of various types of pairs. Because the ANN was trained with only 54 actual pairs with the round-robin method, the 54 pairs may not include the one that was like the test pair. The number of cases used in this study was rather small; however, since only a limited number of this type of observer study has been reported, it is uncertain whether this type of study was feasible and whether some important results could be obtained. Therefore, this study was conducted as a pilot study with a small number of

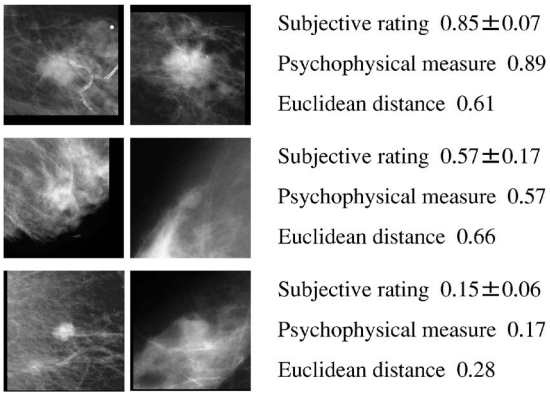


FIG. 10. Examples of pairs of masses which had good agreement between the radiologists' average subjective similarity ratings and psychophysical measures.

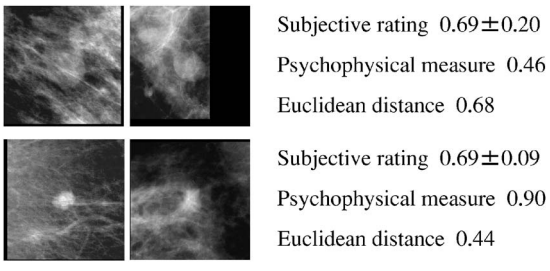


FIG. 11. Examples of pairs of masses which had relatively large differences between the radiologists' average subjective similarity ratings and psychophysical measures.

cases. However, a larger observer study would be needed for determination of the similarity measures in the future.

This study was based on the same concept of the psychophysical similarity measure by use of the ANN and radiologists' subjective similarity ratings used by Li *et al.*[14] The difference is that, in the previous study, the psychophysical measure was applied to the lung nodules in thoracic CT. These are slice images, and the normal structure around the nodules may not interfere with radiologists' judgment about the nodules. On the other hand, in the present study, the psychophysical measure was applied to masses on mammograms, which are projection images, and soft tissue overlapped with the masses may have affected the observers' judgment. Also, the spatial resolution in CT is on the order of a millimeter, whereas the resolution in mammography is a hundred micrometers. Therefore, the determination of similarity ratings could be more complicated for masses on mammograms than that for nodules in CT. In fact, only three features were found useful for determination of similarity measures by both feature-based and ANN-based techniques in the lung nodule study, whereas six and five features were employed for determination of similarity measures by use of the Euclidean distance and the ANN, respectively, in this preliminary study. The average interobserver correlation coefficients on the subjective ratings for the lung nodule study and our mass study was about the same (0.47 and 0.53, respectively). These results indicate that the variation in subjective ratings on similarity between observers may not depend greatly on the imaging modality used. In our study, radiologists' subjective ratings were obtained by use of a continuous rating scale to reduce bias in using discrete numbers. Although one tenth of a number, such as 1.1, 1.2, 1.3, and so on, was allowed in the nodule study, some observers may not have used these numbers effectively. The use of the continuous rating scale might be one of the reasons for a slightly better interobserver correlation in this study than in the lung nodule study. The result of our study shows that the concept of a psychophysical similarity measure can be applied not only to nodules in thoracic CT, but also to masses on mammograms.

The findings in this preliminary study are encouraging. However, there are several limitations to our study. One of the limitations is that only five radiologists, including two breast radiologists, provided the data for determination of radiologists' subjective similarity ratings. For obtaining more reliable subjective ratings, data from a large number of observers, including more breast radiologists, and/or repeated data from the same observer would be necessary. In this study, only 60 pairs of masses were used. To ensure that psychophysical similarity measures are useful for all kinds of masses, a larger number of pairs of masses with various sizes and types would be necessary for determination of subjective ratings by radiologists. In our preliminary study, five image features were selected for training of the ANN. However, it is not known at present whether those features would be sufficient, or whether additional features would be necessary for a larger number of different pairs of masses. Lastly, the digitized images in the database were collected over more than a decade. The image quality may not be close to the current standard, and could be a source of variability in subjective ratings and in determination of features.

V. CONCLUSION

Our psychophysical similarity measure for pairs of mass lesions on mammograms correlated well with the average subjective ratings given by radiologists. We believe that the psychophysical measure would be useful in the selection of masses similar to an unknown case, which may help radiologists in their diagnosis of breast cancer.

ACKNOWLEDGMENTS

This work was supported by USPHS Grant No. CA62625. The authors are grateful to H. Abe, M.D., Ph.D., F. Li, M.D., Ph.D., H. Nishide, M.S., H. Arimura, Ph.D., and H. Takizawa, Ph.D., for valuable discussions, and to the following for their participation in the observer study: C. Sennett, M.D., M. Kral, M.D., C. Zhang, M.D., S. Paquerault, Ph.D., R. Nishikawa, Ph.D., Y. Jiang, Ph.D., M. Giger, Ph.D., L. Yarusso, I. Bonta, M.D., K. Drukker, Ph.D., J. Wilkie, R. Zur, E. Sidky, Ph.D., B. Liu, Ph.D., M. Stern, Y. Peng, and L. Yu. K.D. and R.A.S. are shareholders of R2 Technology, Inc., Los Altos, CA. 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]Electronic mail: chisa@uchicago.edu

[1]Cancer Facts & Figures 2004. American Cancer Society. Available at www.cancer.org/downloads/STT/CAFF_finalPWSecured.pdf.

[2]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).

[3]E. A. Sickles, "Mamographic features of 300 consecutive nonpalpable breast cancers," Am. J. Roentgenol. **146**, 661–663 (1986).

[4]F. M. Hall, "Nonpalpable breast lesions: Recommendations for biopsy based or suspicion of carcinoma at mammography," Radiology **167**, 353–358 (1988).

[5]A. M. Knutzen and J. J. Gisvold, "Likelihood of malignant disease for various categories of mammmographically detected, nonpalpable breast lesions," Mayo Clin. Proc. **68**, 454–460 (1993).

[6]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).

[7]Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, "Analysis of spiculation in the computerized classification of mammographic masses," Med. Phys. **22**, 1569–1579 (1995).

[8]Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and D. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," Acad. Radiol. **5**, 155–168 (1998).

[9]Z. Huo, M. L. Giger, and C. J. Vyborny, "Computerized analysis of multiple-mammographic view: Potential usefulness of special view mammograms in computer-aided diagnosis," IEEE Trans. Med. Imaging **20**, 1285–1292 (2001).

[10]H. P. Chan, B. Sahiner, 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).

[11]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. **25**, 516–526 (1998).

[12]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).

[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]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).

[15]M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer Jr., "Current states of the Digital Database for Screening Mammography," *Digital Mammography* (Kluwer Academic, Dordrecht, 1998), pp. 457–460.

[16]University of South Florida Digital Mammography Home Page. Available at http://marathon.csee.usf.edu/Mammography/Database.html.

[17]C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data," Stat. Med. **17**, 1033–1053 (1998).

[18]M. L. Giger, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields," Med. Phys. **15**, 158–166 (1988).

[19]M. Aoyama, Q. Li, S. Katsuragawa, H. MacMahon, and K. Doi, "Automated computerized scheme for distinction between benign and malignant solitary pulmonary nodules on chest images," Med. Phys. **29**, 701–708 (2002).

[20]M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision* (Brooks/Cole, Pacific Grove, 1999).

[21]D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," *Parallel Distributed Processing* (MIT, Cambridge, MA, 1986).

[22]S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., (Prentice Hall, Upper Saddle River, NJ, 1999).

[23]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).

[24]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).

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

[26]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).

[27]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).

[28]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).