Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules

Qiang Li,[a] Feng Li, Junji Shiraishi, Shigehiko Katsuragawa, Shusuke Sone,[b] and Kunio Doi
Department of Radiology, The University of Chicago, Chicago, Illinois 60637

(Received 4 September 2002; accepted for publication 17 March 2003; published 17 September 2003)

We have been developing a computerized scheme to assist radiologists in improving the diagnostic accuracy for lung cancers on low-dose computed tomography (LDCT) scans by use of similar images for malignant nodules and benign nodules. A database of 415 LDCT scans including 73 cases with 76 confirmed cancers and 342 cases with 413 confirmed benign nodules was first collected in an LDCT screening program for early detection of lung cancers in Nagano, Japan. An observer study by use of receiver operating characteristics analysis was first conducted with five radiologists to demonstrate that presenting similar images for malignant nodules and benign nodules can significantly improve radiologists' performance in the diagnosis of unknown nodules. Another observer study was then conducted for obtaining reliable data on subjective similarity ratings by 10 radiologists. Based on the subjective similarity ratings, three important features were selected from a number of nodule features, and four different techniques for the determination of similarity measures, namely, a feature-based technique, a pixel-value-difference based technique, a cross-correlation-based technique, and a neural-network-based technique, were investigated and evaluated in terms of the correlation coefficient with the subjective similarity ratings. The experimental results in this study indicated that the neural-network-based technique can provide a reliable psychophysical similarity measure which is comparable to the subjective similarity ratings for a single radiologist when evaluated by use of correlation with the average similarity ratings for the other nine radiologists. © *2003 American Association of Physicists in Medicine.*
[DOI: 10.1118/1.1605351]

Key words: lung nodule, low-dose CT, similarity measure, similar nodule, artificial neural network

I. INTRODUCTION

Lung cancer is the leading cause of deaths in the U.S. among all types of cancer.[1] It causes more than 150 000 deaths in the U.S. each year, which is more than the total number of deaths resulting from colon cancer, breast cancer, and prostate cancer combined. Early detection and treatment of lung cancer are effective ways of improving the survival rate, and have been attempted in the U.S. and Japan by use of computed tomography (CT).[2–5] Computer-aided diagnostic (CAD) schemes for nodule detection are effective methods for assisting radiologists in the early detection of lung cancer in thoracic CT scans.[6–12] The current CAD schemes for nodule detection in low-dose CT (LDCT) generally achieved a detection sensitivity of 70%–85% with tens of false positives per case.

It is well-known that distinguishing between malignant and benign lung nodules in CT scans is a difficult task for radiologists,[13–16] and that a vast majority of lung nodules detected in CT screening programs were benign and thus were false positive findings.[2–5] According to recent findings on a lung cancer screening program with LDCT images, 500 (83%) of 605 patients with suspicious pulmonary nodules were proved to have benign lesions, whereas only 105 (17%) patients were confirmed to have malignant nodules.[5] Therefore, a number of research groups have attempted to develop CAD schemes for nodule differentiation in CT images in order to achieve a low false positive (benign nodule) rate.[17–22] These CAD schemes generally achieved an Az value of 0.85–0.95 for the distinction between benign and malignant nodules. It should be noted that these CAD schemes were based on high resolution CT images, probably because LDCT images are generally considered to be inappropriate for diagnosing nodule.

We believe, however, that LDCT images are still useful for nodule diagnosis, if radiologists can confidently eliminate some benign nodules based on LDCT findings, thus to avoid some unnecessary further examinations. In this study, we attempted to improve radiologists' diagnosis accuracy based on LDCT images by presenting a set of images of malignant and benign nodules similar to an unknown new case to be diagnosed. The reason for presenting similar images is based on the fact that radiologists learn diagnostic skills by observing many clinical cases during their training and clinical practice, and their knowledge obtained from visual impression of images with various diseases constitutes the foundation for their diagnosis. In a similar study, Sklansky *et al.* attempted to develop a mapped-database diagnostic system to reduce the number of benign breast lesions recommended for biopsy and the number of misdiagnosed cancers in mammograms.[23] Their system was designed to map the mul-

tidimensional feature vectors representing the unknown lesion and known lesions into a 2D space, to show the 2D space on a computer screen, and thus to help the radiologists manually find confirmed malignant and/or benign ROIs visually similar to the ROI containing an unknown lesion. The similarity measure between the searched similar lesions and the unknown lesion was not evaluated. Content-based image retrieval is another active research field that employs some common ideas with this study.[24–26] However, in the content-based image retrieval technique, two images are considered as being ''similar'' as long as they are in the same category (human portrait, landscape with mountain and beach, and indoor scene, etc.), even though they may differ in many aspects and may not be visually similar at all.

Two fundamental issues related to the concept of similar images are (1) how radiologists perceive subjectively the similarity between two nodules, and (2) how one can determine a reliable similarity measure that would agree well with the subjective similarity according to radiologists' judgment. If the ''similar'' nodules determined by a computerized scheme are not similar to the unknown nodule in terms of radiologists' visual perception, those nodules would not be useful in assisting radiologists in the diagnosis of the unknown nodule. Therefore, we conducted an observer study with ten radiologists to acquire basic data regarding the subjective similarity ratings which may be related to radiologists' visual perception. Based on these experimental data, we investigated the importance of individual image features and the combination of multiple image features, and we assessed several techniques (such as the use of artificial neural networks) for determination of a reliable similarity measure in order to provide a logical and scientific basis for the selection of similar images for malignant and benign nodules. Methods for measuring subjective ratings in general, particularly in image quality, have been developed by Rockette *et al.*[27] and Good *et al.*[28]

II. MATERIALS

From May 1996 to March 1999, 17 892 examinations on 7847 individuals (with an average age of 66 years) were performed as part of an annual low-dose helical CT (LDCT) screening program for early detection of lung cancers in Nagano, Japan.[3–5] There were 7847 initial examinations performed in the first year, and 5025 and 5020 repeat examinations performed in the following two years. Six hundred and five patients were found with 747 suspicious pulmonary nodules ($<$30 mm) in LDCT, among whom 73 patients were confirmed with 76 primary lung cancer by surgery or biopsy, and 342 patients were confirmed with 413 benign nodules by diagnostic CT, two year follow-up examinations, or surgery. The other patients were suspected to have either malignant or benign nodules, although confirmation was not made on these patients.

A mobile unit equipped with a CT scanner (W950SR, Hitachi, Tokyo) was used for scanning the chest with 10 mm collimation and 10 mm reconstruction interval. Each section consisted of 512×512 pixels, with a pixel size of 0.586 mm, and 4096 (12 bits) gray levels in Hounsfield units. The size ranged from 6 mm to 30 mm (average, 13 mm; standard deviation, 5.4 mm) for malignant nodules, and from 3 mm to 30 mm (average, 9 mm; standard deviation, 4.3 mm) for benign nodules. The location of nodules was identified by a chest radiologist for each of the 489 confirmed nodules (76 malignant, and 413 benign), and a region of interest (ROI) of 42×42 mm^2 (72×72 pixels) was then obtained at the center of a nodule. The ROI size of 42×42 mm^2 was empirically determined because it was considered to be large enough to completely contain the nodules employed in this study. When a nodule was observed in multiple sections, we used only one ROI from the section in which the nodule had the largest area. The 489 ROIs with 76 confirmed malignant nodules and 413 confirmed benign nodules constituted the database used in this study.

III. METHODS AND RESULTS

A. Usefulness of similar images in assisting radiologists diagnosing lung nodules in low-dose CT images

In order to verify whether similar images for malignant and benign nodules can assist radiologists in improving their performance in the diagnosis of an unknown nodule in CT scans, we conducted an observer study in which five radiologists rated the likelihood of malignancy for the unknown nodule without and with the similar nodules. We then evaluated the radiologists' performance without and with the aid of similar nodules by use of receiver operating characteristics (ROC) analysis.

1. Methods

We employed a feature-based technique to search for similar malignant and benign nodules with respect to the unknown nodule to be diagnosed. To do so, a nodule was first segmented from background by using a region growing technique[29–31] and a dynamic programming (DP) technique,[32–34] and then three features, i.e., effective diameter, degree of circularity, and contrast, were determined from the segmented nodule. We selected these three features because they are fundamental image features related to the characterization of a lung nodule by radiologists, and also because we had little knowledge as to which features are effective in the determination of a similarity measure when we conducted this observer study. Therefore, the technique for determination of similar nodules described in this section was preliminary, and has been improved significantly, as will be described later. Each of the three features was normalized such that the mean and the standard deviation of the feature for the set of 489 nodules were 0 and 1, respectively. Finally, a similarity measure was defined in the three-dimensional (3D) feature space as the distance between two nodules, i.e.,

$$d^2(f,g)=(|f(1)-g(1)|^2+|f(2)-g(2)|^2+|f(3)-g(3)|^2)/3,$$

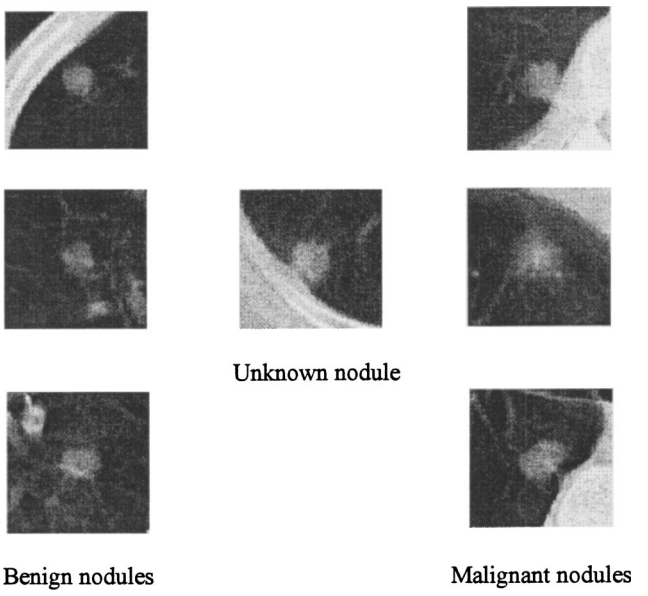

FIG. 1. Illustration for the diagnosis of an unknown nodule with the aid of similar images for three benign nodules and three malignant nodules.

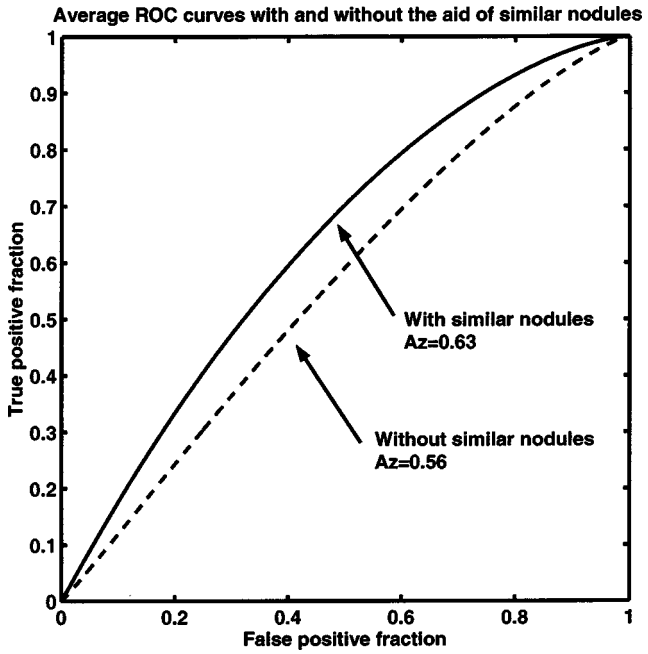

FIG. 2. Comparison of ROC curves for the average performance of the five radiologists in the diagnosis of lung nodules without and with the aid of similar malignant and benign nodules.

where $f=\{f(1),f(2),f(3)\}$ and $g=\{g(1),g(2),g(3)\}$ are the 3D feature vectors representing the two nodules, respectively, and $d(f,g)$ is the similarity measure between the two nodules. The smaller this similarity measure, the more similar the two nodules would be because the features for the two nodules would become similar.

We first randomly selected 36 nodules as unknown ones from the set of 489 nodules (76 malignant, and 413 benign). One half (18) of the unknown nodules were malignant, and the other half (18) were benign. For each of the unknown nodules, we selected the three most similar malignant nodules and the three most similar benign nodules from the remaining 58 malignant nodules and 395 benign nodules, respectively, by use of the above feature-based similarity measure. Five radiologists participated in this observer study, none of whom has viewed the nodules in the database before the study. For each of the unknown nodules, a participating radiologist first rated the likelihood of malignancy based on the observation of the unknown nodule only by marking his/her level of confidence on a line with a continuous rating scale, where the right and left ends of the scale represented definite malignancy and definite benignancy, respectively. Then, the three most similar malignant nodules and the three most similar benign nodules were presented adjacent to the unknown nodule and were shown to the radiologist. Figure 1 illustrates an unknown nodule together with three benign nodules (left-hand side) and three malignant nodules (right-hand side). The radiologist was asked to re-rate the likelihood of malignancy for the unknown nodule after having observed the similar nodules. If the unknown nodule more closely resembles the similar malignant (benign) nodules, it is likely that the radiologist would increase (decrease) the likelihood of malignancy for the unknown nodule. The observer could maintain his/her initial rating if the similar nodules did not provide any new information for his/her judgment. Therefore, for each of the unknown nodules, there were two ratings for the likelihood of malignancy, without and with the aid of similar nodules, respectively. There was no time limit for radiologists to make their decisions.

2. Results

The performance of the five radiologists without and with the aid of similar nodules was evaluated by use of ROC analysis.[35,36] We employed LABMRMC to calculate the average ROC curves and the average Az values (the area under the ROC curve) for the two observation conditions (without or with similar nodules), and to test whether there is significant difference between the two average Az values. Figure 2 shows the average ROC curves for the five radiologists in the diagnosis of lung nodules without and with the aid of similar nodules. The Az value for the average performance of the five radiologists was significantly increased from 0.56 to 0.63 with the aid of similar nodules ($P<0.01$). In fact, all radiologists improved their performance with the aid of similar nodules, and the increase in Az values ranged from 0.05 to 0.12. Therefore, we believe that the radiologists' performance in the diagnosis of lung nodules in CT images can be improved significantly with the aid of similar nodules. It should be noted that the Az values in this observer study were quite low, because the diagnosis of lung nodules in LDCT images is very difficult.

B. Determination of subjective similarity ratings by use of an observer study

1. Methods

We conducted another observer study in order to acquire knowledge concerning the visual perception (or impression)

of similar images by human observers. From this observer study, we wanted to obtain basic data as to how reliable the subjective similarity ratings are, how to improve the reliability of the subjective similarity ratings, and how to utilize the subjective similarity ratings to improve our computerized scheme for evaluation of similar images. We employed the same preliminary technique described above for determination of similar nodules by use of the distance between two 3D feature vectors consisting of effective diameter, degree of circularity, and contrast. Although some of the pairs of nodules selected by the preliminary technique were not similar, dissimilar pairs of nodules were also useful and necessary to be included in this observer study so that a wide range of radiologists' responses on subjective similarities could be included in the data analysis.

We randomly selected 20 ''unknown'' nodules (11 malignant and 9 benign) from the set of 489 nodules, and then determined six ''similar'' malignant and six ''similar'' benign nodules for each ''unknown'' nodule by use of the preliminary technique described above. Therefore, a total of 240 (20×12) pairs of nodules were employed in this observer study. It should be noted that a nodule may be selected as a ''similar'' one for more than once. For example, 20 nodules were selected as similar one for once, 12 nodules for twice, 8 nodules for 3 times, etc. The most frequently selected nodule even appeared as similar one for as many as 10 times. Ten radiologists and 10 physicists participated in the observer study. Each of them rated the subjective similarity independently based on the overall impression for each of the 240 pairs of nodules, with the following rating scores:

0, the two nodules are not similar;

1, the two nodules are somewhat similar;

2, the two nodules are very similar;

3, the two nodules are almost identical.

The observers were allowed to use fractional numbers, such as 1.1, 1.2, or 1.3, to express a similarity rating.

2. Results

We found in this study that there was a large variation among the subjective similarity ratings assessed by individual radiologists. The average correlation coefficient for all pairs of two radiologists among the 10 radiologists was only 0.47. Therefore, it is difficult to obtain reliable subjective similarity ratings from a single radiologist. We also calculated the average correlation coefficient between the similarity ratings of a single radiologist and the average similarity ratings of the other nine radiologists. To do so, a radiologist's similarity ratings were temporarily excluded, and the average similarity ratings for the other nine radiologists were computed. The correlation coefficient between this average similarity ratings and the radiologist's similarity ratings that were temporarily excluded was calculated. This process was repeated 10 times, namely, once for each of the 10 radiologists. The average correlation coefficient for the 10 iterations was calculated to be 0.62, which is significantly higher than the average correlation coefficient value of 0.47 for all pairs of two radiologists. It is apparent that the reliability of the average similarity ratings for nine radiologists was improved compared with that of a single radiologist. In order to illustrate the reliability of the average similarity ratings for all 10 radiologists, Fig. 3 shows the relationship between the average subjective similarity ratings assessed by 10 radiologists and 10 physicists. It is apparent in Fig. 3 that the average subjective similarity ratings assessed by the 10 radiologists correlate well with those assessed by the 10 physicists. The correlation coefficient between the two average similarity ratings was 0.88, which is a remarkably high value compared with that between two radiologists. Therefore, we will employ as ''gold standard'' the average subjective similarity ratings assessed by the 10 radiologists to improve our computerized scheme for the determination of similarity measures. It may be important to note that the subjective judgments on the similarity of lung nodules in LDCT by nonmedically trained human observers (10 physicists) appeared to be highly correlated with and thus very similar to those by radiologists. However, the issues related to subjective judgment by different groups of observers need to be investigated further.

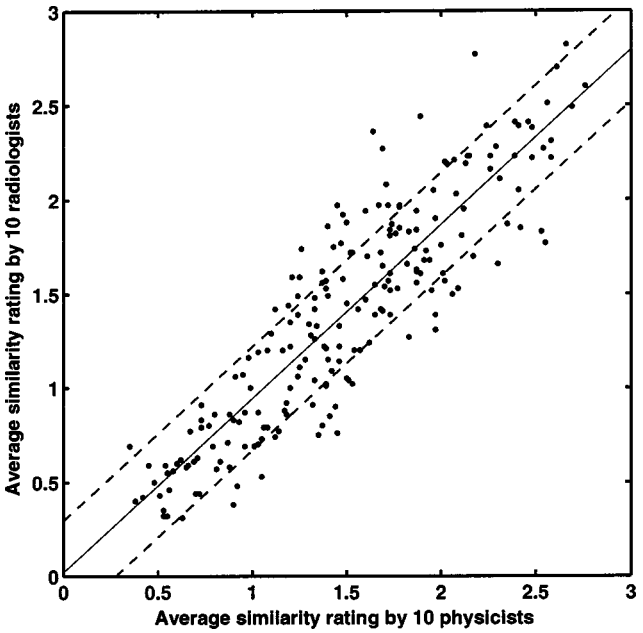

FIG. 3. Relationship between the average subjective similarity ratings assessed by 10 radiologists and 10 physicists. The solid line shows regression line and the two dashed lines indicate one standard deviation from the regression line.

C. Determination of similarity measures

1. Methods

Although an automated technique for nodule segmentation was employed for determination of a similarity measure in the initial observer studies above, it appears that more accurate results for the similarity measures can be obtained by use of the nodule outlines manually delineated by radiologists, because even relatively small errors in nodule seg-

TABLE I. Features employed for the determination of similarity measures.

Feature	Definition	Significance
Effective diameter	Diameter of an "equivalent" circle with the same area as that of the nodule	Malignant nodules have larger diameter value
Degree of circularity	Ratio of the overlap area of the nodule and the equivalent circle to the total area of the nodule	Malignant nodules have smaller circularity value
Degree of irregularity	One minus the ratio of the perimeter of the equivalent circle to that of the nodule	Malignant nodules have larger irregularity value
CT value	Average CT value over a 7×7 region at the center of the nodule	Malignant nodules have smaller CT value
Contrast	Difference in average CT value between the 7×7 region above and the ring-shaped background region	Malignant nodules have smaller contrast value
Pixel standard deviation	Standard deviation of the pixel values over the nodule	Malignant nodules have larger value
Radial gradient index (RGI)	Ratio of average magnitude value of edge gradient projected to the radial direction to that of edge gradient without projection	Malignant nodules have smaller RGI value

mentation seem greatly to affect the accuracy of features and thus the similarity measures. Therefore, a radiologist manually delineated the outline for each of the 489 nodules, which were employed for computation of nodule features hereafter. In addition to the delineated nodule region in the ROI, we also automatically determined a ring-shaped background region with a width of 5 mm which was immediately adjacent to the nodule outline. This nodule background region was employed for calculation of some features such as the contrast.

(a) Determination of similarity measures based on nodule features: Table I shows the definition and the significance of the seven features [effective diameter, degree of circularity, degree of irregularity, CT value, contrast, pixel standard deviation, and radial gradient index (RGI)] employed in this study for the determination of similarity measures. These features were selected because they were considered to be important to radiologists in their distinction between malignant and benign nodules.[37,38] We then determined the Euclidean distance $d(f,g)$ between a pair of nodules in feature space,

$$d^2(f,g)=\frac{1}{N}\left(\sum_{m=1}^{N}|f(m)-g(m)|^2\right), \tag{1}$$

where $f=\{f(1),f(2),\ldots,f(N)\}$ and $g=\{g(1),g(2),\ldots,g(N)\}$ are the N-dimensional feature vectors for the two nodules, respectively. A disadvantage in using the above distance as a similarity measure is its reverse correlation (i.e., negative correlation coefficient) with the subjective similarity rating. To address this problem, we employed the following exponential function for conversion of the distance in the feature space to a similarity measure:

$$s(f,g)=3\times e^{-A\times d(f,g)},$$

where $s(f,g)$ is the similarity measure, $d(f,g)$ is the Euclidean distance in the feature space, and A is a constant to be determined. A scaling factor of 3 was used to adjust the similarity measure in the same range as that for subjective similarity ratings. The constant A was equal to 0.98 in this study; it was determined automatically by use of a least square method[39] for maximizing the correlation coefficient between the similarity measure $s(f,g)$ and the subjective similarity ratings by 10 radiologists for the 240 pairs of nodules. Various combinations of features were tested, and their performance for the determination of similarity measures was compared, as will be described later.

(b) Determination of similarity measures based on the pixel-value-difference technique: The similarity measure defined above is based on the similarity of the features for a pair of nodules. The technique employed in this section is based on the pixel values of the two images to be compared.[40] We first calculate the root mean square (RMS) difference in pixel values between the two nodules in ROIs I and J by the following equation:

$$d^2(I,J)=\frac{1}{|D|}\left(\sum_{(m,n)\ \text{in}\ D}|I(m,n)-J(m,n)|^2\right),$$

where D is the intersection of two regions in the two ROIs, each of which includes the nodule area and the ring-shaped background area; and $|D|$ is the number of pixels inside the region D. We then employed another exponential function to convert the RMS pixel difference into a similarity measure that has a positive correlation coefficient with the subjective similarity rating, i.e.,

$$s(I,J)=3\times e^{-B\times d(I,J)},$$

where $s(I,J)$ is the similarity measure, $d(I,J)$ is the RMS pixel difference, and B is a constant. In this study, the constant B was determined to be 0.008 by use of the least square method[39] for maximizing the correlation coefficient between the similarity measure $s(I,J)$ and the subjective similarity ratings for the 240 pairs of nodules.

(c) Determination of similarity measures based on cross-correlation technique: We also employed a cross-correlation technique for the determination of a similarity measure between two images to be compared. The cross-correlation coefficient was defined by

$$c^2(I,J)=\frac{1}{|D|}\left(\sum_{(m,n)\in D}\frac{\{I(m,n)-\overline{I}\}\{J(m,n)-\overline{J}\}}{\sigma_I\sigma_J}\right),$$

where $c(I,J)$ is the cross-correlation coefficient between the two nodules in ROIs I and J; D is a region defined in the

above section; $|D|$ is the number of pixels inside D; $\overline{I}$ and σ_I are the mean and the standard deviation of the pixel values inside region D of the ROI I, respectively; and $\overline{J}$ and σ_J are the mean and the standard deviation of the pixel values inside region D of the ROI J, respectively. The mean and the standard deviation of the pixel values inside region D of ROIs I and J are defined by the following equations:

$$\overline{I}=\frac{1}{|D|}\left(\sum_{(m,n)\in D} I(m,n)\right), \quad \overline{J}=\frac{1}{|D|}\left(\sum_{(m,n)\in D} J(m,n)\right),$$

$$\sigma_I^2=\frac{1}{|D|}\left(\sum_{(m,n)\in D} |I(m,n)-\overline{I}|^2\right),$$

$$\sigma_J^2=\frac{1}{|D|}\left(\sum_{(m,n)\in D} |J(m,n)-\overline{J}|^2\right).$$

Again, an exponential function was employed to convert the cross-correlation coefficient to a similarity measure whose range is the same as that of subjective similarity ratings,

$$s(I,J)=3\times e^{-C\times(1-c(I,J))},$$

where $s(I,J)$ is the similarity measure, $c(I,J)$ is the cross correlation coefficient, and C is a coefficient of 5.47 determined by use of the least square method[39] for maximizing the correlation coefficient between the similarity measure $s(I,J)$ and the subjective similarity ratings for the 240 pairs of nodules.

(d) Determination of psychophysical similarity measure by use of an artificial neural network: We used an artificial neural network (ANN) for the determination of a psychophysical similarity measure based not only on the objective features and objective measures, but also on the subjective similarity ratings. We employed a three-layer ANN with an input layer, an output layer, and a hidden layer.[37,41,42] The input units represented various objective features/measures determined from a pair of nodules to be compared, and the single output unit represented a new similarity measure for the pair of nodules. In the process of training for the ANN, the subjective similarity ratings were employed as the teaching signal, i.e., the output of the ANN. It should be noted, therefore, that the ANN was trained to learn the relationship between the various objective features/measures of two nodules and the corresponding subjective similarity ratings by radiologists. Thus, once training was completed, the ANN output would provide a psychophysical similarity measure for a given set of objective features/measures which would correlate well with the subjective similarity ratings. In this study, a round-robin (leave-one-out) method was used for verifying the effectiveness of the ANN. With this method, one pair of nodules was excluded from the 240 pairs of nodules, and the remaining 239 pairs were used for training of the ANN. After the ANN was trained, the objective features/measures for the pair of nodules excluded for training were entered as inputs to the ANN for determination of a psychophysical similarity measure. This process was repeated for each of the 240 pairs of nodules one by one, until all psychophysical similarity measures for the 240 pairs of nodules were calculated. Various combinations of objective features/measures for inputs of ANNs were tested, and their performance for the determination of the psychophysical similarity measures were compared, as will be described later.

TABLE II. Correlation coefficients and their 95% confidence intervals between the subjective similarity ratings assessed by 10 radiologists and the computed similarity measures by use of each of the seven features.

Feature used for determination of similarity measures	Correlation coefficient	Confidence interval
Effective diameter	0.48	(0.38,0.57)
CT value	0.35	(0.23,0.46)
Contrast	0.32	(0.20,0.43)
Standard deviation	0.28	(0.16,0.39)
Radial gradient index	0.24	(0.12,0.36)
Degree of circularity	0.24	(0.12,0.36)
Degree of irregularity	−0.02	(−0.15,0.11)

2. Results

In this study, the quality of a computed similarity measure was evaluated by use of the correlation coefficient with the subjective similarity ratings assessed by ten radiologists for the 240 pairs of nodules. The greater the correlation coefficient, the more important the computed similarity measure in the determination of similar images. For the feature-based method, we first attempted to evaluate the importance of each feature for the determination of similarity measures. Table II lists the correlation coefficients and their 95% confidence intervals[43] between the subjective similarity ratings and the feature-based similarity measures by use of each of the seven features. It is apparent that nodule size (effective diameter), nodule contrast (contrast and CT value) provide moderate correlation values with the subjective similarity measures by 10 radiologists. Pixel value variation over a nodule (pixel standard deviation and radial gradient index) and the degree of circularity provide relatively weak correlations with the subjective similarity measures. It should be noted that the degree of irregularity, which is generally considered to be important and frequently employed for the distinction between malignant nodules and benign nodules, does not seem to have correlation with the subjective similarity ratings by 10 radiologists.

For evaluating the importance of the combinations of multiple features, the feature-based similarity measures were calculated for multiple features. We found that the combination of effective diameter and CT value provided a good result among all possible combinations of two features; the correlation coefficient between the similarity measure and the similarity rating was 0.57. The combination of effective diameter, CT value, and RGI provided another good result (correlation coefficient of 0.60) among all possible combinations of three features. We also investigated the similarity measures by use of more than three features, and we found that their benefits were either negligible or decreased compared with the use of the combination of the effective diam-

TABLE III. Correlation coefficients and their 95% confidence intervals between the subjective similarity ratings by 10 radiologists and the computed similarity measures by use of feature-based, pixel-value-difference-based, cross-correlation-based, and ANN-based techniques.

Techniques	Correlation coefficient	Confidence interval
Feature-based (effective diameter, CT value, and radial gradient index)	0.60	(0.51,0.68)
Pixel-value-difference-based	0.49	(0.39,0.58)
Cross-correlation-based	0.45	(0.34,0.55)
ANN-based (7 inputs, i.e., diameter, CT value, and RGI for two nodules, and pixel difference)	0.72	(0.65,0.78)

eter, CT value, and RGI. Therefore, we used these three features in our computerized scheme for determination of similar images.

Table III lists the correlation coefficients and their 95% confidence intervals[43] between the subjective similarity ratings and the computed similarity measures by use of the feature-based, the pixel-value-difference-based, the cross-correlation-based, and the ANN-based techniques. It is apparent that the feature-based technique (0.60) provided superior result to the pixel-value-difference-based (0.49) and the cross-correlation-based techniques (0.45). However, the psychophysical similarity measure determined by use of the ANN-based technique provided the highest correlation coefficient (0.72) among the techniques that we investigated.

Various combinations of objective features/measures for inputs of ANNs were tested for the determination of the psychophysical similarity measures. Table IV shows the correlation coefficients and their 95% confidence intervals[43] between the subjective similarity ratings and psychophysical similarity measures obtained with different combinations of objective features/measures. It should be noted that the three features (effective diameter, CT value, and RGI) used in the ANNs were first selected based on their high correlation with the subjective similarity ratings. For the first three ANNs in Table IV, the inputs of the ANNs included (a) six features (three from each of the two nodules to be compared), (b) three differences in features between the two nodules, and (c) the combination of the six features and the three differences. The psychophysical similarity measure using the differences in the three features alone provided a deteriorated correlation coefficient (0.60) compared with that (0.68) using the six features determined from the two nodules; this is understandable because the three differences did not provide all of the information included in the six feature values. However, the correlation coefficient (0.64) obtained with the psychophysical similarity measure using the combination of the six features and the three differences was also lower than that obtained using the six features.

TABLE IV. Correlation coefficients and their 95% confidence intervals between the subjective similarity ratings assessed by 10 radiologists and the psychophysical similarity measures obtained with various combinations of objective measures.

Inputs to ANN for determination of similarity measure	Correlation coefficient	Confidence interval
(a) Six inputs (diameter, CT value, and RGI for two nodules)	0.68	(0.61,0.74)
(b) Three inputs (difference in diameter, CT value, and RGI between two nodules)	0.60	(0.51,0.68)
(c) Nine inputs (diameter, CT value and RGI for two nodules and their difference)	0.64	(0.56,0.71)
(d) Seven inputs (diameter, CT value, and RGI for two nodules, and pixel difference)	0.72	(0.65,0.78)
(e) Seven inputs (diameter, CT value, and RGI for two nodules, and cross correlation)	0.64	(0.56,0.71)

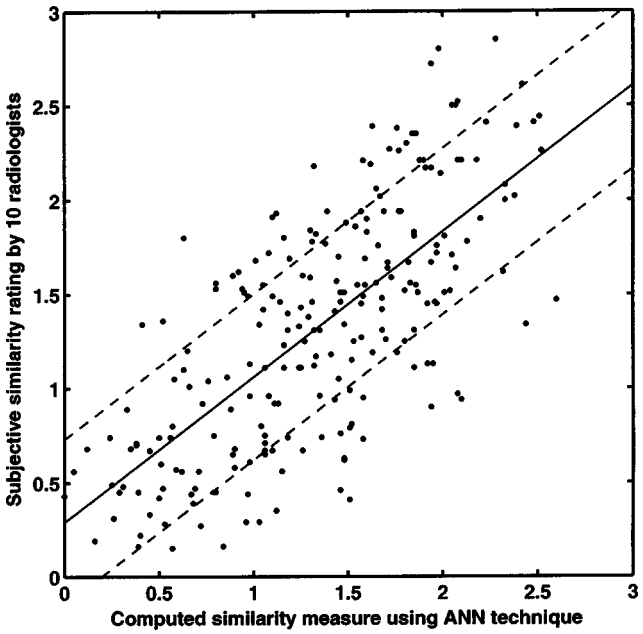

FIG. 4. Relationship between the psychophysical similarity measures based on an ANN and the subjective similarity ratings assessed by 10 radiologists. The ANN has seven input units, including six features and the pixel-value difference between two nodules. The solid line shows regression line and the two dashed lines indicate one standard deviation from the regression line.

We also examined the combination of the six features with objective measures obtained with (d) the pixel-value-difference technique or (e) the cross-correlation technique. The psychophysical similarity measure obtained with the combination of the six features and the pixel-value-difference value provided a relatively large correlation coefficient (0.72) with the subjective similarity rating by 10 radiologists. However, the inclusion of the cross-correlation value did not improve the correlation coefficient (0.64). Figure 4 shows the relationship between the psychophysical similarity measures obtained with the ANN of (d) and the subjective similarity ratings by ten radiologists for the 240 pairs of nodules. Because the average correlation coefficient between the similarity ratings by a single radiologist and the average similarity ratings by the other nine radiologists was only 0.62, the psychophysical similarity measures determined by use of the ANN seems to provide a high accuracy that is comparable to the subjective similarity ratings obtained by a single radiologist.

Finally, it is important to note that the LDCT images em-

ployed in this study were obtained with a single scanner; therefore, the image quality depended on the scanner and the specific reconstruction algorithm used. The readers should realize that we did not take the scanner dependency into account in the search algorithm for similar nodules.

IV. DISCUSSION

Similarity measure plays an important role in many applications of pattern recognition and computer vision, such as image matching and registration, cluster analysis,[40] face detection and recognition,[44] and content-based image database retrieval.[24–26] There are many kinds of definitions for the similarity measure, among which are various forms of distance (Euclidean distance, Minkowsky distance, Mahalanobis distance, and Hausdorff distance[45]), cross-correlation, and mutual information measures.[46] Distance-based similarity measures are widely used in many applications, and can be applied in both feature space and image space; for example, the pixel-value-difference technique in this study employed a distance in image space. Cross correlation is also well known for its simplicity in implementation and its relatively long computation time, and is applied mainly in image space. However, in most applications, the evaluation of similarity measures for a specific application has been generally ignored. Therefore, it is often unclear why one selects one specific similarity measure over another. To address this problem, in this study, the average subjective similarity ratings were obtained with 10 radiologists in advance, and were employed as gold standard to evaluate the effectiveness of four similarity measures used in the search of subjectively similar images. Moreover, the subjective similarity ratings were utilized to provide a new psychophysical similarity measure which correlates well with radiologists' subjective ratings.

Content-based image retrieval has been a very active research field in recent years. It is motivated by the fact that it is difficult to efficiently manage, browse, search, and retrieve a multimedia database with a huge amount of image and video information.[24–26] On the one hand, the content-based image retrieval technique is applied in an analogous way as the similar nodule searching technique is used in this study; namely, given a query image, it tries to search a number of images that contain similar image content as does the query image. In addition, from a technical point of view, extraction of features (such as color, texture, and shape) and definition of similarity measures are common issues in content-based image retrieval as well as in our similar nodule searching technique. On the other hand, however, there is an important difference between the two similar image searching techniques. In the content-based image retrieval technique, images are classified into a number of categories according to their contents, such as human portrait, landscape with mountain and beach, and indoor scene. Two images are considered as being ''similar'' as long as they are in the same category, even though they may differ in many aspects and may not be visually similar. In this study, however, all images contained the same object (namely, a nodule), and any pair of images can be considered as being ''similar'' in terms of the definition of similarity used in the content-based image retrieval technique. Therefore, a stricter definition of similarity based on the overall visual impression was employed in this study, which represents a more challenging problem than that in the content-based image retrieval technique.

We have previously developed a computerized scheme to help radiologists improve their performance in the diagnosis of nodules in chest radiographs, in which the likelihood of malignancy for a lesion was presented to radiologists.[37,47] Although the likelihood of malignancy for lung nodules would be an important aid to radiologists, the numeral alone might not be adequate and convincing enough for radiologists. In this study, we showed that similar images can provide radiologists with visual aid, and would be useful for improving radiologists' performance in the diagnosis of nodules. Similar images may also be combined with the likelihood of malignancy to further improve radiologists' diagnostic accuracy. In addition, similar images can be applied to the diagnosis of various lesions in images obtained with different modalities, and used to help radiologists learn the diagnosis of very difficult and complex abnormalities, such as the differential diagnosis of interstitial lung diseases in chest images.

V. CONCLUSION

The similar images for malignant nodules and benign nodules have potential to significantly improve radiologists' performance in the diagnosis of unknown nodules. The average subjective similarity ratings obtained by ten radiologists are important for the selection of features and for the evaluation of different techniques for calculating similarity measures A psychophysical similarity measure can be determined based on the use of ANN with objective features/measures and subjective rating data. The psychophysical similarity measure on pairs of CT nodules provided more reliable results compared with objective similarity measures.

ACKNOWLEDGMENTS

This work was supported by USPHS Grant No. CA62625. The authors are grateful to E. Lanzl for improving the paper, to K. Suzuki and Y. Uchiyama for helpful discussion, and to the following radiologists for their participation in the observer test conducted in this study: H. MacMahon, U. Bick, C. Vyborny, H. Abe, P. MacEneaney, B. O'Rourke, B. Ward, S. Regalado, and E. O'Conner. K.D. is a shareholder of R2 Technology, Inc., Los Altos, CA, and Deus Technologies, Inc., Rockville, MD. 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]Author to whom correspondence should be addressed; electronic mail: qiangli@uchicago.edu
[b]Also at: JA Azumi General Hospital, 3207-1 Ikeda, Nagano 399-8695 Japan.

[1] S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo, ''Cancer statistics 2000,'' Ca-Cancer J. Clin. **50**, 7–33 (2000).

[2] C. I. Henschke, D. I. McCauley, D. F. Yankelovitz, D. P. Naidich, G. McGuinness, O. S. Miettinen, D. M. Libby, M. W. Pasmantier, J. Koizumi, N. K. Altorki, and J. P. Smith, "Early lung cancer action project: overall design and findings from baseline screening," Lancet **354**, 99–105 (1999).
[3] S. Sone, S. Takashima, F. Li, Z. Yang, T. Honda, Y. Maruyama, M. Hasegawa, T. Yamanda, K. Kubo, K. Hanamura, and K. Asakura, "Mass screening for lung cancer with mobile spiral computed tomography scanner," Lancet **351**, 242–245 (1998).
[4] S. Sone, F. Li, Z. Yang, S. Takashima, Y. Maruyama, M. Hasegawa, J. Wang, S. Kawakami, and T. Honda, "Characteristics of small lung cancers invisible on conventional chest radiography: analysis of 44 lung cancers detected by population-based screening using low-dose spiral CT," Br. J. Radiol. **73**, 137–145 (2000).
[5] S. Sone, F. Li, Z. Yang, T. Honda, Y. Maruyama, S. Takashima, M. Hasegawa, S. Kawakami, K. Kubo, M. Haniuda, and T. Yamada, "Results of three-year mass screening programme for lung cancer using mobile low-dose spiral computed tomography scanner," Br. J. Cancer **84**, 25–32 (2001).
[6] T. Okumura, T. Miwa, J. Kako, S. Yamamoto, M. Matsumoto, Y. Tateno, T. Iinuma, and T. Matsumoto, "Variable N-Quoit filter applied for automatic detection of lung cancer by x-ray CT," in *Computer-Assisted Radiology*, edited by H. U. Lemke, M. W. Vannier, K. Inamura, and A. Farman (Elsevier Science, New York, 1998), pp. 242–247.
[7] Y. Lee, T. Hara, H. Fujita, S. Itoh, and T. Ishigaki, "Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique," IEEE Trans. Med. Imaging **20**, 595–604 (2001).
[8] K. Kanazawa, Y. Kawata, N. Niki, H. Satoh, H. Ohmatsu, R. Kakinuma, M. Kaneko, K. Eguchi, and N. Moriyama, "Computer aided diagnostic system for pulmonary nodules based on helical CT images," in *Computer-Aided Diagnosis in Medical Imaging*, edited by K. Doi, H. MacMahon, M. L. Giger, and K. R. Hoffmann (Elsevier Science, New York, 1999), pp. 131–136.
[9] M. S. Brown, M. F. McNitt–Gary, J. G. Goldin, R. D. Suh, J. W. Sayre, and D. R. Aberle, "Patient-specific models for lung nodule detection and surveillance in CT images," IEEE Trans. Med. Imaging **20**, 1242–1250 (2001).
[10] M. N. Gurcan, B. Sahiner, N. Petrick, H.-P. Chan, E. A. Kazerooni, P. N. Cascade, and L. Hadjiiski, "Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system," Med. Phys. **29**, 2552–2558 (2002).
[11] A. G. Armato, M. L. Giger, C. Moran, J. T. Blackburn, K. Doi, and H. MacMahon, "Computerized detection of pulmonary nodules on CT scans," Radiographics **19**, 1303–1311 (1999).
[12] S. G. Armato, M. L. Giger, and H. MacMahon, "Automated lung segmentation in digitized posteroanterio chest radiographs," Acad. Radiol. **5**, 245–255 (1998).
[13] S. S. Siegelman, N. F. Khouri, F. P. Leo, E. K. Fishman, R. M. Braverman, and E. A. Zerhouni, "Solitary pulmonary nodules: CT assessment," Radiology **160**, 319–327 (1986).
[14] C. V. Zwirewich, S. Vedal, and R. R. Miffler, "Solitary pulmonary nodule: high-resolution CT and radiologic correlation," Radiology **179**, 469–476 (1991).
[15] M. Gurnay and S. J. Swensen, "Solitary pulmonary nodules: determining the likelihood of malignancy with neural network analysis," Radiology **196**, 823–929 (1995).
[16] Y. Matsuki, K. Nakamura, H. Watanabe, T. Aoki, H. Nakata, S. Katsuragawa, and K. Doi, "Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: evaluation with receiver operating characteristic analysis," AJR, Am. J. Roentgenol. **178**, 657–663 (2002).
[17] D. F. Yankelevitz *et al.*, "Small pulmonary nodules: Volumetrically determined growth rates based on CT evaluation," Radiology **217**, 251–256 (2000).
[18] M. F. McNitt-Gray *et al.*, "A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: Preliminary results," Med. Phys. **26**, 880–888 (1999).
[19] N. Wyckoff *et al.*, "Classification of solitary pulmonary nodules (SPNs) imaged on high resolution CT using contrast enhancement and three dimensional quantitative image features," Proc. SPIE **3979**, 1107–1115 (2000).
[20] M. F. McNitt-Gray *et al.*, "Computer-aided diagnosis of the solitary pulmonary nodule imaged on CT:2D, 3D and contrast enhancement features," Proc. SPIE **4322**, 1845–1852 (2001).
[21] Y. Kawata *et al.*, "Classification of pulmonary nodules in thin section CT images based on shape characterization," Proc. ICIP **3**, 528–531 (1997).
[22] Y. Kawata *et al.*, "Computer aided differential diagnosis of pulmonary nodules using carvature based analysis," Proc. ICIAP 470–475 (1999).
[23] 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 microcalcification," Acad. Radiol. **7**, 395–405 (2000).
[24] M. Flickner *et al.*, "The QBIC project: querying images by content using color, texture and shape," *SPIE Proceedings of Storage and Retrieval of Image and Video Databases* (Bellingham, WA, 1993), pp. 173–181.
[25] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: content-based manipulation of image databases," *SPIE Proceedings of Storage and Retrieval of Image and Video Databases II* (Bellingham, WA, 1994), pp. 34–47.
[26] Y. Rui, T. S. Huang, and S. F. Chang, "Image retrieval: Current techniques, promising direction, and open issues," J. Visual Commun. Image Represent **10**, 39–62 (1999).
[27] H. E. Rockette, C. M. Johns, J. L. Weissman, J. M. Holbert, J. H. Sumkin, J. L. King, and D. Gur, "Relationship of subjective ratings of image quality and observer performance," Proc. SPIE **3036**, 152–159 (1997).
[28] W. F. Good, G. Maitz, J. L. King, R. Gennari, and D. Gur, "Observer performance assessment of JPEG-compressed high-resolution chest images," Proc. SPIE **3663**, 8–13 (1999).
[29] 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).
[30] M. L. Giger, K. Doi, H. MacMahon, C. Metz, and F. F. Yin, "Pulmonary nodules: Computer-aided detection in digital chest images," Radiographics **10**, 41–51 (1990).
[31] X. Xu, K. Doi, T. Kobayashi, H. MacMahon, and M. L. Giger, "Development of an improved CAD scheme for automated detection of lung nodules in digital chest images," Med. Phys. **24**, 1395–1403 (1997).
[32] M. Aoyama, Q. Li, S. Katsuragawa, F. Li, S. Sone, and K. Doi, "Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images," Med. Phys. **30**, 387–394 (2003).
[33] A. A. Amini, T. E. Weymouth, and R. C. Jain, "Using dynamic programming for solving variational problem in vision," IEEE Trans. Pattern Anal. Mach. Intell. **12**, 855–867 (1990).
[34] H. Yamada, C. Merritt, and T. Kasvand, "Recognition of kidney glomerulus by dynamic programming matching method," IEEE Trans. Pattern Anal. Mach. Intell. **10**, 731–737 (1988).
[35] C. E. Metz, "ROC methodology in radiologic imaging," Invest. Radiol. **21**, 720–733 (1986).
[36] C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," Invest. Radiol. **24**, 234–245 (1989).
[37] K. Nakamura, H. Yoshida, R. Engelmann, H. MacMahon, S. Katsuragawa, T. Ishida, K. Ashizawa, and K. Doi, "Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks," Radiology **214**, 823–830 (2000).
[38] Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schimidt, "Analysis of spiculation in the computerized classification of mammographic masses," Med. Phys. **22**, 1569–1579 (1995).
[39] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, Cambridge, MA, 1986), pp. 498–546.
[40] A. K. Jain, *Fundamentals of Digital Image Processing* (Prentice-Hall, New York, 1989), pp. 400–402.
[41] K. Ashizawa, T. Ishida, H. MacMahon, C. Vyborny, S. Katsuragawa, and K. Doi, "Artificial neural networks in chest radiography: Application to the differential diagnosis of interstitial lung diseases," Acad. Radiol. **6**, 2–9 (1999).
[42] Z. Huo, M. L. Giger, and C. E. Metz, "Effect of dominant features on neural network performance in the classification of mammographic lesions," Phys. Med. Biol. **44**, 2579–2595 (1999).
[43] D. G. Altman, *Practical Statistics in Medical Research* (CRC, Boca Raton, FL, 1990), pp. 293–294.

[44] E. Hjelmas and B. K. Low, ''Face detection: A survey,'' Comput. Vis. Image Underst. **83**, 236–274 (2001).
[45] D. P. Huttenlocher, G. A. Klanderman, and W. J. Ruckidge, ''Comparing images using Hausdorff distance,'' IEEE Trans. Pattern Anal. Mach. Intell. **15**, 50–863 (1993).
[46] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, ''Multi-modality image registration by maximization of mutual information,'' IEEE Trans. Med. Imaging **16**, 187–198 (1997).
[47] M. Aoyama, Q. Li, S. Katsuragawa, and K. Doi, ''Automated computerized scheme for distinction between benign and malignant solitary pulmonary nodules on chest images,'' Med. Phys. **29**, 701–708 (2002).