Journal of Digital Imaging

Assessment of Performance and Reliability of Computer-Aided Detection Scheme Using Content-Based Image Retrieval Approach and Limited Reference Database

Xiao Hui Wang,[1] Sang Cheol Park,[1] and Bin Zheng[1]

Content-based image retrieval approach was used in our computer-aided detection (CAD) schemes for breast cancer detection with mammography. In this study, we assessed CAD performance and reliability using a reference database including 1500 positive (breast mass) regions of interest (ROIs) and1500 normal ROIs. To test the relationship between CAD performance and the similarity level between the queried ROI and the retrieved ROIs, we applied a set of similarity thresholds to the retrieved similar ROIs selected by the CAD schemes for all queried suspicious regions, and used only the ROIs that were above the threshold for assessing CAD performance at each threshold level. Using the leave-one-out testing method, we computed areas under receiver operating characteristic (ROC) curves (A_Z) to assess CAD performance. The experimental results showed that as threshold increase, (1) less true positive ROIs can be referenced in the database than normal ROIs and (2) the A_Z value was monotonically increased from 0.854±0.004 to 0.932 ±0.016. This study suggests that (1) in order to more accurately detect and diagnose subtle masses, a large and diverse database is required, and (2) assessing the reliability of the decision scores based on the similarity measurement is important in application of the CBIR-based CAD schemes when the limited database is used.

KEY WORDS: Content-based image retrieval, computer-aided diagnosis (CAD), cancer detection, computerized method

INTRODUCTION

In the medical imaging research field, a large number of computer-aided detection and diagnosis (CAD) schemes of medical images have been developed and extensively assessed. Studies have demonstrated the significant potential of using CAD to help improve performance and efficiency of radiologists in reading and interpreting medical images to detect and diagnose suspicious abnormalities[1,2]. However, the limitation of using the traditional (or "black-box" type) CAD schemes in the clinical practice was also well recognized[3,4]. Due to its advantages and potential of being used as a "visual aid" tool, the content-based image retrieval (CBIR) approach has been recently applied in CAD development and applications in an attempt to increase radiologists' confidence in accepting CAD-cued results and, thus, improve detection/diagnostic performance[5,6]. In particular, developing CBIR-based CAD schemes of mammograms (for detecting breast masses and micro-calcification clusters) has been attracting extensive research interest in the last decade. Several such CBIR-based CAD schemes have been developed and tested in previous studies[7–14]. Although a number of different image search or retrieval methods have been applied and used in CBIR approaches, the decision indices or detection scores of the CAD schemes are commonly computed by the differently weighted ratios between the "most similar" true-positive (TP) and false-positive (FP) images or regions of interests (ROIs) selected (retrieved) by the CBIR algorithms from the pre-established reference databases. As a result, the quality (including the size and diversity) of the reference databases plays an important role in developing CAD schemes using CBIR approach[15].

[1]*From the Department of Radiology, University of Pittsburgh, 3362 Fifth Avenue, Pittsburgh, PA, 15213, USA.*

Correspondence to: Xiao Hui Wang, Department of Radiology, University of Pittsburgh, 3362 Fifth Avenue, Pittsburgh, PA, 15213, USA; e-mail: xwang@mail.magee.edu

Online publication 5 March 2010

doi: 10.1007/s10278-010-9281-x

In the CBIR-based CAD schemes, the reference images (or ROIs) depicting true-positive or false-positive lesions (normal tissue structures) are typically selected from the limited available medical image databases. The number of selected ROIs in the reference databases ranged from 57[10] to 3,000 ROIs[13] in the previously reported studies. The most of the previous studies used a leave-one-out validation method to test the performance of the CAD schemes in which assuming there were *N* ROIs in the reference database, the CAD scheme queried (tested) each ROI in the reference database once and the CBIR algorithm searched through the rest of the database (excluding the queried one) to identify the *K* reference ROIs that are considered “the most similar” to the queried ROI and then computed a detection score (the likelihood of this queried ROI being true-positive). Once the *N* detection scores were generated for all *N* ROIs in the reference database, the researchers assessed and reported CAD performance by computing the area under a receiver operating characteristic (ROC) curve (A_Z value). However, by forcing CAD schemes to compute detection scores for all queried ROIs without assessing the actual similarity levels between the queried ROIs and the retrieved similar reference ROIs, this type of CBIR application and performance assessment method ignores two important issues related to the effectiveness of the reference database. First, due to diversity of abnormalities (lesions) depicted on medical images, the limited number of reference ROIs can only be very sparsely distributed in relationship to the complex image feature space[16]. Since a limited image database cannot be adequate to cover the whole image feature space, the CBIR algorithms can retrieve reference ROIs that have very high level of similarity to some queried ROIs but relatively lower level of similarity to the others. Second, the available references in the database are typically not uniformly distributed in the feature space. Some regions in the feature space include higher concentrated clusters of ROIs, and the other regions only have sparsely distributed reference ROIs. Therefore, the impact of using the limited reference databases on the overall performance of the CBIR-based CAD scheme and the reliability of individual query results have not been fully investigated in previous studies. Our hypotheses of this study are that (1) the similarity levels between the queried ROIs and the retrieved similar reference ROIs varies widely using the limited reference databases when applying the CBIR algorithms to a set of diverse testing images and (2) the similarity level can be an important index to assess the reliability of CAD-generated decision indices (detection scores) for both clinical relevance and visual similarity. If these hypotheses can be validated in this study, we can then provide our suggestions about (1) how to develop the optimal CBIR-based CAD schemes and/or (2) how to more objectively evaluate the performance and reliability of CBIR-based CAD schemes using the limited reference databases.

MATERIALS AND METHODS

To validate our hypotheses, we first assembled a relatively large reference database of ROIs depicting suspicious breast masses in this study and then conducted an experiment to systematically remove the queried ROIs that have relatively lower similarity scores computed by a multi-image feature-based CBIR algorithm. We investigated the relationship between the similarity scores of the queried ROIs and the performance of a CBIR-based CAD scheme. This experiment aims to help researchers better understand how to optimally assess the adequacy or diversity of the reference database and what is the impact of the current practice to force CAD scheme computing a decision index (a detection score) without considering the actual similarity level between the queried ROI and the retrieved reference ROIs.

Because of the nature of the study (a blinded retrospective study of anonymized cases after removal of protected health information and without identifiers), there are no Health Insurance Portability and Accountability Act concerns associated with this study. The examinations were acquired under our Institutional Review Board (IRB) approved protocols and permission to use the examinations for studies had been obtained.

The detailed descriptions of our database, the applied CBIR approach, and the evaluation of experimental procedures are reported here.

Reference Database

We have built a large and diverse database of digitized mammograms in our research laboratory. The original mammograms were initially generated using several film digitizers with the pixel size of 50×50 μm and 12-bit gray level resolution. The digitized images were then subsampled by the factor of 2 (increasing the pixel size to 100× 100 μm) and saved in our database. The reference database selected in this study includes 3,000 ROIs extracted from this pre-established digitized mammogram database in our laboratory. Each ROI has a fixed size of 512×512 pixels. Among the 3,000 selected ROIs in the reference database, 1,500 true-positive ROIs depict either pathology verified malignant masses (1,290) or biopsy proved benign masses (210). The remaining 1,500 negative ROIs were extracted from the image areas depicting normal breast tissue but were falsely detected as masses by the CAD scheme previously developed in our research laboratory[17]. Thus, all suspicious mass regions (including both true-positive and false-positive ones) were initially segmented and detected by the CAD scheme. Since a fraction of incorrectly segmented mass regions by the CAD scheme, which is unavoidable in a large and diverse image database, could substantially affect the accuracy of computed image features and eventually CAD classification performance[18], the boundary contours of these automatically segmented mass regions were visually examined and manually corrected (if needed). The detailed description of this CAD-based mass segmentation procedure along with the number of mass regions whose boundary contours have been manually corrected in this reference database has been reported in our previous study[19]. In brief, the automatically segmented boundary contours in 19.2% (288 out of 1,500) mass regions showed noticeable error and were manually corrected in this database.

After mass region segmentation, we used a computer scheme to compute 14 morphological and intensity distribution features from each ROI. These include three global (whole breast area)-based image features namely, (1) average pixel value in breast area, (2) average, and (3) standard deviation of local pixel value fluctuation in the breast area. The other 11 image features are region-based local features that are computed from a rectangular frame that covers the segmented suspicious mass area plus the extension of 25 pixels in all four directions. These 11 ROI-based features are (1) region conspicuity, (2) normalized mean radial length of the region, (3) standard deviation of radial length, (4) skew of radial length, (5) shape factor ratio of the region, (6) standard deviation of pixel value inside the mass region, (7) standard deviation of gradient of boundary pixels, (8) skew of gradient of boundary pixels, (9) standard deviation of pixel value in the surrounding background, (10) average local pixel value fluctuation in the surrounding background, and (11) normalized central position shift between the region center pixel and the pixel with minimum digital value inside the region. The detailed definitions and computing methods of these 14 image features have also been reported in our previous study[20]. All feature values were normalized to be distributed between 0 and 1. These 14 normalized image features were then saved into a feature data file with all extracted ROIs in our reference database.

A CBIR Scheme

We applied a CBIR scheme that uses a multi-feature-based *k*-nearest neighbor (KNN) algorithm to search for the similar breast masses depicted on the reference database. Once a testing ROI is queried, the CBIR scheme searches for *K* most similar ROIs (i.e., $K=15$) from the reference database. This KNN-based CBIR scheme has been previously optimized, using genetic algorithm, and reported[13,20]. In brief, the similarity is measured by the difference in feature values, $d(q,r_i)$, between a queried ROI ($f_q(x_i)$) and a reference ROIs ($f_r(x_i)$) in a multidimensional ($n=14$) feature space.

$$d(q,r_i)=\sqrt{\sum_{i=1}^{n}\left(f_q(x_i)-f_r(x_i)\right)^2}$$

A similarity score (index) to measure the similarity level between the queried ROI and each of the retrieved reference ROI is defined as:

$$S(q,r_i)=\frac{1}{d(q,r_i)^2}$$

As a result, the smaller feature difference ("distance") generates the larger similarity score indicating the high level of similarity between the queried ROI and the retrieved reference ROI. To detect the likelihood of the queried ROI depicting a true-positive mass based on the comparison of *K* most similar reference ROIs, we applied and tested two types of decision indices (detection scores) that have been commonly used in different CBIR-based CAD schemes in this study. By assuming that the K most similar reference ROIs include N true-positive ROIs and M false-positive ROIs $(N + M = K)$, we defined and computed the first decision index as:

$$D_1(q) = \frac{\sum_{i=1}^{N} S\left(q, r_i^{\mathrm{TP}}\right)}{\sum_{i=1}^{N} S(q, r_i^{\mathrm{TP}}) + \sum_{j=1}^{M} S\left(q, r_j^{\mathrm{FP}}\right)}.$$

The second decision index is computed as:

$$D_2(q) = \frac{1}{K}\sum_{i=1}^{N} S\left(q, r_i^{\mathrm{TP}}\right) - \frac{1}{K}\sum_{j=1}^{M} S\left(q, r_j^{\mathrm{FP}}\right).$$

Evaluation Method

To test the performance of this CBIR based CAD scheme, we applied a leave-one-out validation method. In this testing method, each of 3,000 ROIs was selected as a queried ROI once and CBIR scheme searched for the K=15 most similar reference ROIs from the rest of 2,999 ROIs (excluding itself) stored in the reference database. For each test, the scheme computed the similarity scores ($S(q,r_i)$, i=1,2,...,K) and two decision indices or detection scores ($D_1(q)$ and $D_2(q)$) that indicate the likelihood of the queried ROI depicting a true-positive mass. After 3,000 iterations, two sets of detection scores were generated for all 1,500 true-positive ROIs and 1,500 false-positive ROIs. We then applied a ROC data fitting and analysis program (ROCKIT[21]) to compute two ROC curves including the areas under ROC curves (A_Z values) and their standard deviations using two sets of detection scores. The A_Z value is used as an index to assess the performance of the CBIR-based CAD scheme in selecting clinically relevant reference ROIs.

To investigate the relationship between CAD performance including its reliability and the similarity scores of the retrieved ROIs to the queried ROIs, we first sorted all 3,000 ROIs based on the largest similarity score $(\mathrm{MAX}|S(q, r_i)|,\ i = 1, 2, \ldots, K)$ of each ROI. We computed the average (μ) and standard deviation (σ) of the largest similarity scores among these 3,000 ROIs. The interval $(\mu - 2\sigma, \mu + 2\sigma)$ of similarity scores was normalized to the range between 0 and 1. All similarity scores falling outside the interval range were assigned to the nearest ending normalized value. For example, if $S(q, r_i)\mu - 2\sigma$, its normalized score is assigned to 0, otherwise, if $S(q, r_i)\mu + 2\sigma$, its normalized score is assigned to 1. After sorting these 3,000 ROIs, we applied a set of nine sequential threshold values namely from 0.1 to 0.9 at an interval of 0.1 to the largest normalized similarity scores of all queried ROIs. In each threshold, all ROIs whose normalized similarity scores are smaller than the threshold were removed, which means that no single retrieved reference ROI is similar to the queried ROI based on this threshold value of the similarity score. The ROC program was applied to the remaining ROIs to compute A_Z value and reassess CAD performance using a new subset of queried ROIs with higher similarity level (scores). The experimental results were then compared and analyzed in this study.

RESULTS

Two histograms (plotted in Fig. 1) demonstrate the distribution between the number of TP or FP ROIs and threshold values on the normalized similarity scores. Although the average value and standard deviation of the largest normalized similarity scores of 3,000 queried ROIs is 0.384±0.308, there is an obvious difference in the similarity scores between

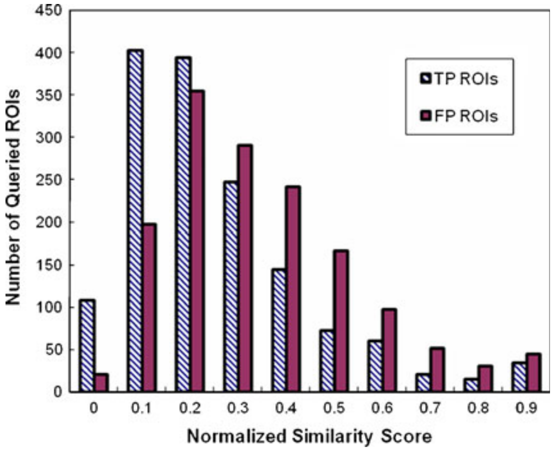

Fig. 1. Comparison of histograms of normalized similarity scores between 1500 positive ROIs that depict verified masses (TP ROIs) and 1,500 negative ROIs that depict CAD-cued false-positive masses (FP ROIs). It shows that diversity level of TP ROIs is larger than FP ROIs.

Table 1. Number of Queried True-Positive (TP) and False-Positive (FP) ROIs with the Largest Similarity Scores Greater than Threshold Value on the Normalized Similarity Scores

Threshold	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TP ROIs	1,500	1,392	989	595	348	203	130	70	49	34
FP ROIs	1,500	1,479	1,281	926	635	393	226	128	76	45

It shows that feature variation in TP regions is greater than FP regions. As a result, when applying the sequential thresholds (from 0.1 to 0.9) on the normalized similarity scores to remove the queried ROIs with lower similarity scores, more TP ROIs than FP ROIs are removed

TP and FP ROIs in this reference database. The average values and standard deviations of the largest normalized similarity scores of 1,500 TP and 1,500 FP ROIs are 0.345±0.339 and 0.425±0.253, respectively. The average similarity score of TP ROIs is lower than that of FP ROIs, which indicates that the diversity level of image features among the TP breast masses is bigger than that of the suspicious but negative breast tissue regions selected in this reference database. As a result, when applying the sequential threshold values on the normalized similarity scores to remove the queried ROIs with lower similarity scores, more TP ROIs than FP ROIs were removed (Table 1).

The experimental results also demonstrate that by systematically identifying and removing the queried ROIs that have lower level of similarity to the CBIR scheme-selected (retrieved) reference ROIs, we substantially increased the performance of CAD scheme in detecting true-positive breast masses in this study. For example, when using the first decision index ($D_1(q)$) the original CAD performance level (A_i value) using all 3,000 ROIs is 0.854±0.004. When assessing CAD performance using a subset of 34 TP ROIs and 45 FP ROIs that have the highest similarity level to the CBIR scheme-selected reference ROIs (threshold=0.9), the A_Z value increases to 0.932±0.016. Although the CAD scheme performance levels (the A_Z values) using the two decision indices ($D_1(q)$ and $D_2(q)$) tested in this study are slightly different (Table 2), the trend of monotonic increase of the A_Z values as the increase of threshold values on the normalized similarity scores remains quite similar for the use of both decision indices (Figs. 2 and 3). Table 3 summarizes the statistical results of applying the least-square regression method to fit the linear relationship between the threshold values of the similarity scores and the computed A_Z values of CAD performance. The significance of the increasing trend (the p value) of the linear relationship was computed using the ANOVA test. In summary, the results indicate that the higher similarity level (score) between the queried ROI and the CBIR-retrieved reference ROIs, the higher accuracy and reliability of the final CAD-generated detection score on this queried ROI is.

DISCUSSION

In the clinical practice of interpreting medical images, radiologists routinely refer to and compare the similar cases with verified results in their decision making of detecting and diagnosing suspicious lesions. As the advance of digital imaging technologies, using computerized CBIR approaches has shown significant advantages to assist radiologists in interpreting digital medical images[5,22]. When a CBIR-based CAD scheme is used as a "visual aid" tool[7,20], radiologists' confidence in their decision making to consider and/or accept the CAD-generated likelihood score of a suspicious lesion being true-positive (or malignant) depends on a number of facts including but not limited to (1) the performance of the CAD schemes and (2) whether the selected "most

Table 2. Areas Under ROC Curves (A_Z Values) Using Two Different Decision Indices as the Threshold Values on the Normalized Similarity Scores Change from 0 to 0.9

Threshold	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$D_1(q)$	0.854	0.859	0.859	0.864	0.877	0.888	0.908	0.911	0.919	0.932
$D_2(q)$	0.838	0.840	0.849	0.858	0.868	0.866	0.881	0.882	0.887	0.898

It shows that the trend of monotonic increase of the A_Z values as the increase of thresholds on the normalized similarity scores remains quite similar for the use of both decision indices

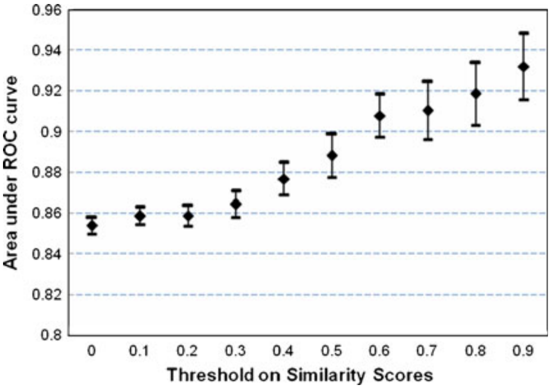

Fig. 2. The change of CBIR scheme performance (A_Z values) using the decision index $D_1(q)$ as the increase of threshold values on the similarity scores of the queried ROIs. It shows a trend of monotonic increase of the A_Z values as the increase of thresholds on the normalized similarity scores using decision index $D_1(q)$.

similar" ROIs by CBIR scheme are actually visually similar to the queried ROI. This study clearly demonstrated that due to the diversity of medical images (in particular for the true-positive lesions) and the use of the limited available reference database, it is extremely difficult to identify and select a set of reference ROIs that can maintain the very comparably similar levels to all queried (testing) ROIs (in particular the subtle true-positive lesions). Thus, if the CAD system displays a set of the CBIR-selected "most similar" ROIs that have actually lower similarity scores, radiologists are likely to ignore the CAD-cued results for these queried ROIs and reduce their overall confidence in the CAD system including the cuing results for other queried ROIs[23]. We emphasize that the number of the most similar reference ROIs selected by CBIR schemes for generating CAD scores (i.e., 15 in this study) is often different from the actual number of similar reference ROIs used as "visual aid" in attempt to achieve the balance between information provided to the radiologists and their workload (reading efficiency). In previously reported studies, the number of ROIs actually showed to the radiologists limited from 6[23] to 12[7].

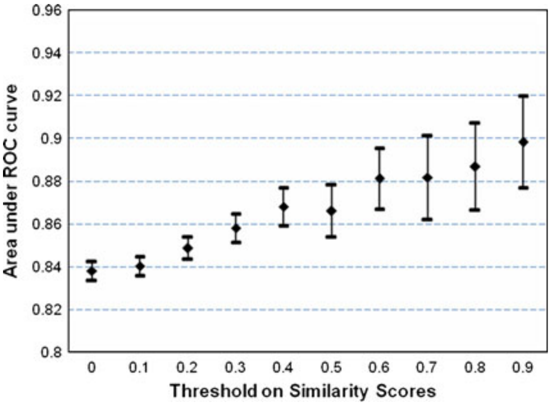

Fig. 3. The change of CBIR scheme performance (A_Z values) using the decision index $D_2(q)$ as the increase of threshold values on the similarity scores of the queried ROIs. It shows a trend of monotonic increase of the A_Z values as the increase of thresholds on the normalized similarity scores using decision index $D_2(q)$.

Table 3. The Statistic Results of the Linear Regression Between the Threshold Values on the Similarity Scores and the A_Z Values of ROC Curves Generated by CAD Scheme Using Two Decision Indices

Decision index	*R* square	Standard error	*P* value
$D_1(q)$	0.9594	0.0061	7.53E-7
$D_2(q)$	0.9769	0.0033	7.79E-8

The significance of the increasing trend (the *p* value) of the linear relationship was demonstrated using the ANOVA test

The results of this study support the previous finding that because as a local databased machine learning and optimization approach, the CBIR-based CAD scheme was more sensitive to the quality of training (reference) database than the other CAD schemes optimized using a global databased machine learning approach (i.e., the artificial neural network), the CBIR-based CAD scheme achieved significantly lower performance without considering the actual similarity between the queried ROIs and the selected reference ROIs[24]. Therefore, blandly forcing the CBIR-based CAD scheme to compute the likelihood (detection) score of the queried ROI being true-positive without considering the actual similarity level between the queried ROIs and the retrieved reference ROIs may often generate unreliable results for both clinical relevance and visual similarity (if the CAD is used as an "visual aid" tool). To solve this problem, we recommend that when applying CBIR-based CAD schemes, one should either report both of the CBIR-generated similarity score and CAD-generated detection score for each queried ROI or skip (discard) any queried ROIs that have lower similarity scores by reporting them as undecided (or unclassified) ROIs.

Unlike some of previous studies in developing CBIR-based CAD schemes in which the negative (false-positive) ROIs were randomly selected from the negative mammograms, each of negative ROIs selected in our reference database depicts one growth area detected and segmented by CAD scheme as suspicious regions (false-positive). As a result, the same set of image features can be computed for both true-positive and false-positive ROIs. We found in this experiment that in applying each similarity threshold, higher percentage of true-positive ROIs than false-positive (negative) ROIs was eliminated, which means that it is generally more difficult for TP ROIs to find highly similar reference ROIs in the limited available reference database. This indicates that the true-positive breast masses usually have much larger diversity. As a result, increase of the number of true-positive ROIs may be more important than increase of negative (false-positive) ROIs in establishment of the reference databases used in CBIR-based CAD schemes. In addition, the ROC analysis results show that as the size of reference database reduces, the uncertainty (standard deviation of A_Z values) also increases (Figs. 2 and 3) indicating that building and using a large and diverse reference database can also increase the reliability of evaluating CAD performance.

In summary, we investigated and assessed the relationship between the performance of the CBIR-based CAD scheme and the use of the limited reference database in this study. Unlike some previous studies that suggested that the optimal approach to develop CBIR-based CAD schemes was to use the database with intelligently selected small number of reference ROIs (i.e., 10 to 20[25]), this study demonstrated that without a large and diverse reference database the overall performance of CBIR-based CAD scheme was substantially reduced and many queried regions (in particular the subtle ones) might not find the similar reference regions resulting in the reduction of the reliability of CAD-generated likelihood (detection) scores of being malignant for these queried regions. Based on the results of this study, we propose two recommendations in developing and evaluating the CBIR-based CAD schemes. First, in order to more accurately and reliably detect and diagnose subtle breast masses (or other types of abnormalities), one needs to continuously build a large and diverse reference database that makes the selected reference ROIs be more uniformly distributed in the image feature space. Second, when only a limited database is available during the preliminary CAD development, one needs to make the scheme enable to monitor and report the index (score) of the similarity levels between the queried ROI and the retrieved reference ROIs to improve the performance and reliability of the final CAD-cueing results. If CAD is used as a "visual aid" tool, providing the similarity score along with the CAD-generated detection score for each queried ROI may also minimize the risk of misleading the radiologists and increase their confidence in CAD-cued results. Such a hypothesis needs to be further investigated in the future studies.

ACKNOWLEDGEMENTS

This work is supported in part by Grants 1 UL1 RR024153 from the National Center for Research Resources and CA77850, CA101733 from the National Cancer Institute, National Institutes of Health, to the University of Pittsburgh.

REFERENCES

1. Nishikawa RM: Current status and future directions of computer-aided diagnosis in mammography. Comput Med Imaging Graph 31:224–235, 2007

2. Hirose T, Nitta N, Shiraishi J, et al: Evaluation of computer-aided diagnosis (CAD) software for the detection of lung nodules on multidetector row computed tomography (MDCT): JAFROC study for the improvement in radiologists' diagnostic accuracy. Acad Radiol 15:1505–1512, 2008

3. Gur D, Sumkin JH, Rockette HE, et al: Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. J Natl Cancer Inst 96:185–190, 2004

4. Nishikawa RM, Kallergi M: Computer-aided detection in its present form is not an effective aid for screening mammography. Med Phys 33:811–814, 2006

5. Muller H, Michoux N, Bandon D, Geissbuhler A: A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. Int J Med Informatics 73:1–23, 2004

6. Zheng B: Computer-aided diagnosis in mammography using content-based image retrieval approaches: current status and future perspectives. Algorithm 2:828–849, 2009

7. Giger ML, Huo Z, Vyborny CJ, et al: Intelligent CAD workstation for breast imaging using similarity to known lesions and multiple visual prompt aides. Proc SPIE 4684:768–773, 2002

8. El-Naga I, Yang Y, Galatsanos NP, et al: A similarity learning approach to content-based image retrieval: application to digital mammography. IEEE Trans Med Imaging 23:1233–1244, 2004

9. Wei C, Li C, Wilson R: A general framework for content-based medical image retrieval with its application to mammograms. Proc SPIE 5748:134–143, 2005

10. Alto H, Rangayyan RM, Desautels JE: Content-based retrieval and analysis of mammographic masses. J Electron Imaging 14:023016, 2005

11. Tourassi GD, Harrawood B, Singh S, et al: Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms. Med Phys 34:140–150, 2007

12. Tao Y, Lo SB, Freedman MT, Xuan J: A preliminary study of content-based mammographic masses retrieval. Proc SPIE 6514:65141Z, 2007

13. Zheng B, Mello-Thoms C, Wang X, et al: Interactive computer aided diagnosis of breast masses: computerized selection of visually similar image sets from a reference library. Acad Radiol 14:917–927, 2007

14. Rosa NA, Felipe JC, Traina AJ, et al: Using relevance feedback to reduce the semantic gap in content-based image retrieval of mammographic masses. Conf Proc IEEE Med Biol Soc 2008:406–409, 2008

15. Park SC, Sukthankar R, Mummert L, et al: Optimization of reference library used in content-based medical image retrieval scheme. Med Phys 34:4331–4339, 2007

16. Zheng B, Chang YH, Good WF, Gur D: Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme. Acad Radiol 4:497–502, 1997

17. Gur D, Stalder JS, Hardesty LA, et al: Computer-aided detection performance in mammographic examination of masses: assessment. Radiology 233:418–423, 2004

18. Zheng B, Pu J, Park SC, Zuley M, Gur D: Assessment of the relationship between lesion segmentation accuracy and computer-aided diagnosis scheme performance. Proc SPIE 6915:691530-1–691530-11, 2007

19. Wang X, Park SC, Zheng B: Improving performance of content-based image retrieval schemes in searching for similar breast mass regions: an assessment. Phys Med Biol 54:949–961, 2009

20. Zheng B, Lu A, Hardesty LA, et al: A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment. Med Phys 33:111–117, 2006

21. Metz CE: *ROCKIT 0.9B Beta version,* University of Chicago, http://www-radiology.uchicago.edu/krl/KRL_ROC/software_index6.htm, 1998.

22. Lehmann TM, Guld MO, Deselaers T, et al: Automatic categorization of medical images for content-based retrieval and data mining. Comput Med Imaging Graph 29:143–155, 2005

23. Zheng B, Abrams G, Britton CA, et al: Evaluation of an interactive computer-aided diagnosis system for mammography: a pilot study. Proc SPIE 6515:65151M-1–65151M-8, 2007

24. Park SC, Pu J, Zheng B: Improving performance of computer-aided detection scheme by combining results from two machine learning classifiers. Acad Radiol 16:266–274, 2009

25. Mazurowski MA, Zurada JM, Tourassi GD: Selection of samples in case-based computer-aided decision systems. Phys Med Biol 53:6079–6096, 2008