Optimization of reference library used in content-based medical image retrieval scheme

Sang Cheol Park
Department of Radiology, University of Pittsburgh, 3362 Fifth Avenue, Pittsburgh, Pennsylvania 15213

Rahul Sukthankar and Lily Mummert
Intel Research Pittsburgh, 4720 Forbes Avenue, Pittsburgh, Pennsylvania 15213

Mahadev Satyanarayanan
School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213

Bin Zheng[a]
Department of Radiology, University of Pittsburgh, 3362 Fifth Avenue, Pittsburgh, Pennsylvania 15213

Building an optimal image reference library is a critical step in developing the interactive computer-aided detection and diagnosis (I-CAD) systems of medical images using content-based image retrieval (CBIR) schemes. In this study, the authors conducted two experiments to investigate (1) the relationship between I-CAD performance and size of reference library and (2) a new reference selection strategy to optimize the library and improve I-CAD performance. The authors assembled a reference library that includes 3153 regions of interest (ROI) depicting either malignant masses (1592) or CAD-cued false-positive regions (1561) and an independent testing data set including 200 masses and 200 false-positive regions. A CBIR scheme using a distance-weighted K-nearest neighbor algorithm is applied to retrieve references that are considered similar to the testing sample from the library. The area under receiver operating characteristic curve (A_z) is used as an index to evaluate the I-CAD performance. In the first experiment, the authors systematically increased reference library size and tested I-CAD performance. The result indicates that scheme performance improves initially from $A_z = 0.715$ to 0.874 and then plateaus when the library size reaches approximately half of its maximum capacity. In the second experiment, based on the hypothesis that a ROI should be removed if it performs poorly compared to a group of similar ROIs in a large and diverse reference library, the authors applied a new strategy to identify "poorly effective" references. By removing 174 identified ROIs from the reference library, I-CAD performance significantly increases to $A_z = 0.914$ ($p < 0.01$). The study demonstrates that increasing reference library size and removing poorly effective references can significantly improve I-CAD performance. © *2007 American Association of Physicists in Medicine.* [DOI: 10.1118/1.2795826]

I. INTRODUCTION

In the clinical practice of medical image reading and interpretation, radiologists often refer to and compare the similar cases with previously verified results in their decision making of detection and diagnosis. Clinically similar cases can be retrieved from radiologists' brains (previous experience) and available image databases. Advances in digital technologies for computing, networking, and database storage have enabled the automated searching for clinically relevant and visually similar references in large medical image databases. The development of content-based image retrieval (CBIR) technology and schemes has therefore attracted wide research interest in medical imaging research areas.[1,2] For example, to assist radiologists reading and interpreting screening mammograms, researchers have made significant progress during the last 20 years in developing computer-aided detection and diagnosis (CAD) schemes to detect and classify breast abnormalities (i.e., masses and microcalcification clusters) depicted on digitized and digital mammograms.[3] Current commercialized CAD systems cue (highlight) CAD-identified suspicious masses and microcalcification clusters on the images. Although using CAD improves radiologists' efficiency in searching for and detecting microcalcification clusters and may help radiologists detect more cancers associated with malignant microcalcifications,[4] the majority of CAD-cued false-negative cancers associated with malignant masses are discarded by radiologists as false-positives in the clinical environment[5,6] because of (1) the relatively low performance of CAD schemes in mass detection (i.e., higher false-positive rate) and (2) the inability of current systems to explain the reasoning of the CAD decision

making. As a result, the clinical benefit of using current commercialized CAD systems is still under debate and test.[7,8]

To increase radiologists' confidence in CAD-cued suspicious breast masses depicted on mammograms, a number of interactive computer-aided detection and diagnosis (I-CAD) schemes and workstations have been recently developed, tested, and reported by several research groups.[9–16] An I-CAD system aims to provide interactive capability between radiologists and CAD schemes and it includes a reference image library with a set of regions of interest (ROI) depicting verified cases (either malignant lesions or benign and CAD-cued false-positive lesions). Once the radiologist queries a region that depicts a suspicious breast lesion, the I-CAD scheme extracts (or segment) the suspicious lesion and generates a likelihood (detection) score of the queried region depicting a true-positive (or malignant) lesion based on the comparison with a set of similar lesions that are automatically retrieved from the reference library. Then, the detection score and the similar reference regions are displayed in the I-CAD workstation to provide radiologists with a "visual aid" to support clinical decision making. For this purpose, different CBIR schemes have been investigated and tested in developing I-CAD systems. For examples, to search for similar reference mass regions, some have employed the pixel value based information-theoretic similarity measures (e.g., mutual information),[13] while others have used distance (e.g., Euclidean distance) weighted similarity measures based on a multiple-feature space.[14,16] Since the performance of I-CAD systems using CBIR schemes should be assessed using both detection or classification accuracy (selecting clinically relevant reference ROIs) and radiologists' confidence in accepting I-CAD results in their decision making (selecting visually-similar ROIs), some research groups have focused on improving scheme performance measured by the area under receiver operating characteristic (ROC) curve in classification between true-positive and false-positive lesions,[13,14] while others have investigated methods to improve and assess visual similarity of the selected reference regions to the queried lesions.[15,16]

The reference library is one of the most important components in a CBIR-based I-CAD system. In the previous studies, the sizes of reference libraries varied widely from 57 (Ref. 12) to 3000 suspicious mass regions.[16] Although a recent study reported that using an entropy-based index could reduce the size of reference library and computational cost of a mutual information based CBIR scheme while maintaining overall CAD performance,[17] a number of important issues related to the optimization of reference library and its impact on I-CAD performance have not currently been fully investigated. In this study we assembled a relatively large and diverse image reference library as well as an independent testing data set with pathology-verified breast masses and CAD-cued false-positive mass regions. Then we conducted two experiments. The first one investigates the relationship between the increase of reference library size and I-CAD performance in classification between true-positive and false-positive masses. The second experiment develops and tests a new reference selection strategy (method) for building an optimal reference library (i.e., by identifying and removing a small fraction of "poorly effective" or "regionally misfitted" reference samples) and tested the potential improvement of I-CAD performance. These two experiments aim to help us better understand (1) whether using a randomly selected small reference library can achieve comparable testing performance of I-CAD scheme as using a randomly selected large reference library, (2) whether arbitrarily increasing reference library size (unconditionally accepting new labeled samples) can keep improving I-CAD performance, and (3) whether identifying and removing poorly effective ROIs from the reference library can achieve improved and robust results when applying I-CAD scheme to the independent testing data sets.

II. MATERIALS AND METHODS

II.A. A reference library and a testing data set

From a large and diverse image database of digitized mammograms that has been previously established in the Imaging Research Center, Department of Radiology, University of Pittsburgh, we extracted 3553 ROIs that depict either pathology-verified malignant masses (1792) and CAD-cued false-positive mass regions (1761) in this study. The basic image characteristics (including the distribution of mass size and subjectively rated subtleness) of our mammographic image database have been reported in our previous studies.[18,19] For each verified mass region, a computer scheme including the combination of a multilayer topographic region growth algorithm[20] and an active contour algorithm[16] is applied to segment and define the mass region boundary contour. The automated segmentation result is visually examined by the experienced observers and manually corrected if an obvious error of automated segmentation is observed. The boundary contours of all CAD-cued false-positive mass regions are automatically detected by the CAD scheme without any manual modification. After region segmentation, the computer scheme automatically computes and generates a feature vector that includes a set of 36 morphological and intensity-distribution related image features to represent each selected suspicious mass region (including either a true-positive mass or a CAD-cued false-positive region). In this feature vector, nine features are computed from the whole breast area depicted on one digitized mammogram ("global" features) and the remaining 27 features are computed from the segmented mass region and its surrounding background tissue ("local" features). We then computed the mean (μ) and the standard deviation (σ) of each feature from all 3553 selected suspicious mass regions in the image data set. The interval $[\mu - 3\sigma, \mu + 3\sigma]$ of each feature is normalized between 0 and 1. Any feature values falling outside the interval range (outliers) are assigned to either 0 ($< \mu - 3\sigma$) or 1 ($> \mu + 3\sigma$). A detailed description (including the definition and computational method for each of the features) has been reported elsewhere.[19]

From the initially selected 3553 ROIs in the image database, we randomly selected 400 ROIs to create an independent testing data set. Of these 400 ROIs, 200 ROIs depict

TABLE I. The size of five reference libraries used in experiment one.

Library	1	2	3	4	5
Number of true-positive ROIs	318	637	955	1274	1592
Number of false-positive ROIs	312	625	936	1249	1561

true-positive mass regions and 200 depict CAD-cued false-positive mass regions. The remaining 3153 ROIs are used to establish a reference library that includes 1592 ROIs depicting true-positive masses and 1561 ROIs involving CAD-cued false-positive regions.

II.B. A computer scheme to search for similar reference images

In our CBIR scheme a set of image features is used to represent image content. In our previous studies several distance weighted K-nearest neighbor (KNN) algorithms have been investigated and tested to search for similar mass regions from the reference library, which include the KNN algorithms based on weighted Euclidean distance[19] and nonlinear learned distance metrics.[21] In this study we used the CBIR scheme with the linear distance-weighted KNN algorithm to investigate the relationship between the quality of reference library and I-CAD performance.

Using the KNN algorithm, similarity is measured by the distance (d) between a queried mass region (y_q) and each of reference regions (x_i) in a multidimensional (n) feature (f) space

$$d(y_q, x_i) = \sqrt{\sum_{r=1}^{n} (f_r(y_q) - f_r(x_i))^2}. \tag{1}$$

A smaller distance indicates a higher degree of "similarity" between two compared regions. The algorithm selects the K regions that are most similar to the queried mass region from the reference library. A detection score (representing the probability of the queried region being a true-positive mass) is computed as

$$P_{TP} = \frac{\sum_{i=1}^{N} w_i^{TP}}{\sum_{i=1}^{N} w_i^{TP} + \sum_{j=1}^{M} w_j^{FP}}, \tag{2}$$

where $w_i = 1/d(y_q, x_i)^2$ (a distance weight), w_i^{TP} and w_j^{FP} are the distance weight for the true-positive (i) and false-positive (j) mass region, respectively, N is the number of verified true-positive (TP) mass regions, M is the number of CAD-cued false-positive (FP) regions, and $N+M=K$.

In our previous study we have applied a genetic algorithm (GA) to define an optimal topology for the KNN algorithm including the selection of a specific set of features and an optimal number of reference regions (neighbors). From the initial feature pool of 36 global and local image features and using the area under ROC curve (A_z value) as the fitness criterion of GA (or the classification performance index to assess KNN performance), the GA selects a set of 14 image features ($n=14$) and 15 nearest neighbors ($K=15$) to build an optimal KNN algorithm based CBIR scheme. The detailed

description of GA optimization process and the list of 14 selected features have been reported in our previous study.[19] In order to improve visual similarity between the queried mass region and KNN-selected reference regions, three boundary conditions on the difference of region size (area), circularity, and boundary margin spiculation level between a queried region (A_q, C_q, and S_q) and a reference region (A_r, C_r, and S_r) are also implemented in the KNN algorithm, which are: (1) $|A_r - A_q|/A_q \leq 1/3$, (2) $|C_r - C_q| \leq 0.15$, and (3) $S_q = S_r$. As a result, this KNN-based CBIR scheme is restricted to select "similar" regions, each of which has a reasonably comparable size, an overall similar shape, and the same computed boundary spiculation level.[16] Using these three conditions, our two previous observer preference studies demonstrate that visual similarity of automated selected reference regions is significantly ($p < 0.01$) improved over the scheme without these restrictions in searching for similar reference regions.[16,19]

II.C. The relationship between reference library size and I-CAD performance

We conducted two experiments in this study. The first one aims to evaluate I-CAD performance as a function of the number of available reference regions (the size of reference library). In this experiment, we first separately and randomly divided 1592 true-positive and 1561 false-positive mass regions in the original reference library into five exclusive partitions. Each partition has approximately the same number of reference regions (i.e., 318 true-positive mass regions and 312 false-positive ROIs). We then investigated and tested the I-CAD performance by systematically increasing the size of reference library from using two partitions (one for true-positives and one for false-positives) to all ten partitions (five for true-positives and five for false-positives). As a result, a total of five reference libraries are built for this experiment (Table I). For each of the five steps to increase the size of reference library, 400 suspicious mass regions (200 true-positive and 200 false-positive) from an independent testing data set are used as query (testing) mass regions to assess I-CAD performance. For each testing mass region, a detection score is computed by the I-CAD scheme using Eq. (2) as described above. Using the detection scores for all 400 testing regions as a summary index, we computed and plotted the ROC curve using the publicly available ROCFIT software.[22] The area under the ROC curve (A_z value) is used as a performance index to assess I-CAD performance in classifying between true-positive and false-positive mass regions (the selection of clinically relevant and visually similar reference ROIs). In this experiment, five A_z values and the corresponding standard deviations are computed as we increase

the size of the reference library from two partitions (630 ROIs) to ten partitions (3153 ROIs). We plot the distribution diagram between the size of reference library (x axis) and the computed A_z values (y axis). As the size of reference library increases, the statistically significant difference (two-tailed p value) between two adjacent A_z values is also computed.

To investigate whether the number of features used in the CBIR scheme affects the relationship between the I-CAD performance and the increase of reference library size, we repeated this experiment three times by selecting 8, 10, and 12 features from the original 14 features. In each experiment, we iteratively selected a subset of specific features (e.g., 8) until the maximum I-CAD performance is reached when using all available 3153 reference ROIs. We then computed I-CAD performance (A_z values) by changing reference library size. The results of performance changes are tabulated and compared.

II.D. Improving I-CAD performance by building a more effective reference library

Since including redundant and noisy reference samples (e.g., ROIs with low entropy values[17]) in the image reference library can both increase the computational cost and reduce I-CAD performance in classification between true-positive and false-positive masses, the second experiment of this study aims to develop a new reference selection strategy (method) to identify the poorly effective reference ROIs and remove them from the reference library. Then we investigated and tested whether I-CAD performance could be significantly improved by using the "optimal" reference library and the same independent testing data set.

Our hypothesis for identifying a poorly effective reference ROI is that the local neighborhood of such a ROI is typically inconsistent with its label. In other words, a poorly effective ROI tends to superficially resemble ROIs of the wrong class within a large and diverse reference library. Hence, these are regionally misfitted ROIs. Based on this hypothesis, we used the original reference library including all available 3153 ROIs and employed a leave-one-out cross-validation method[23] to identify and remove poorly effective ROIs. Specifically, we selected each ROI stored in the reference library, in turn, as a queried region and searched for similar reference ROIs among the remaining 3152 ROIs in the reference library (excluding the queried region itself) and computed the detection score of this queried ROI (the likelihood of the ROI depicting a true-positive mass). We repeated this process for all of 3153 ROIs in the reference library. Then, we set up two thresholds on the detection scores to identify poorly effective ROIs (one for true-positive and one for false-positive mass regions). These two thresholds are not correlated and can be independently selected. A queried ROI that depicts a true-positive mass is considered a poorly effective ROI if its detection score is smaller than the first threshold, while a poorly effective false-positive ROI is identified if its detection score is larger than the second threshold. These identified poorly effective ROIs are generally surr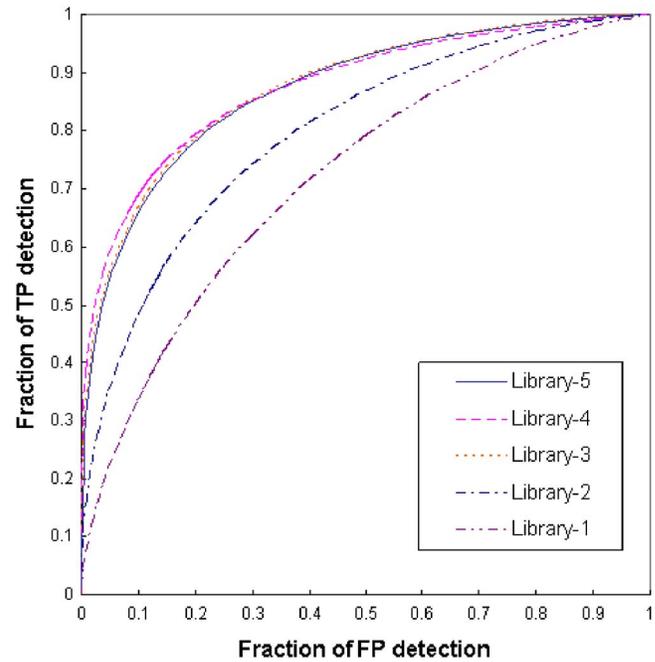ounded by the opposite class of ROIs (i.e., the most of similar neighbors of a "true-positive" ROI are actual false-positive ROIs). We then investigated the selection of these two threshold values to identify and remove the poorly effective ROIs from the reference library and the trend of changing these threshold values to the I-CAD performance. We also compared the difference between I-CAD performance achieved in this experiment and the "maximum" performance obtained in the first experiment.

FIG. 1. Demonstration of five ROC curves of the testing data set using five reference libraries as shown in Table I.
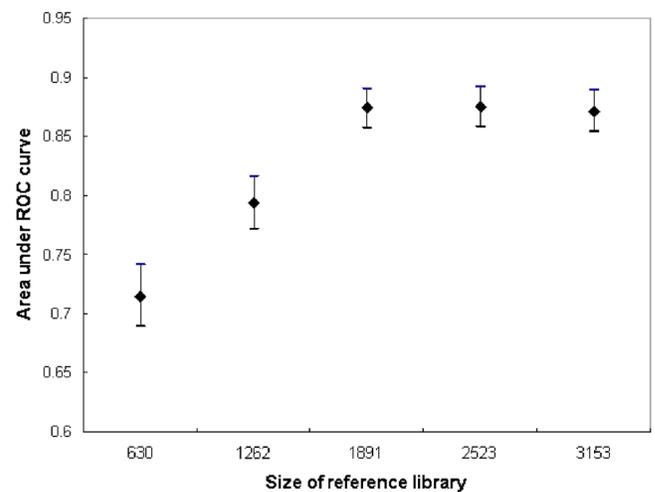

FIG. 2. The trend of the change of the areas under ROC curve (A_z values) as the increase of reference library size. The corresponding computed A_z values (from the left to the right) are 0.715±0.026, 0.794±0.023, 0.874±0.017, 0.875±0.017, and 0.872±0.017, respectively.

TABLE II. Change of I-CAD performance (A_z values) as the change of reference library size and number of features used in KNN. (Note: The standard deviation of A_z values varies from ±0.015 to ±0.026.)

Number of features	Library 1	Library 2	Library 3	Library 4	Library 5
8	0.686	0.725	0.836	0.835	0.835
10	0.701	0.767	0.845	0.847	0.849
12	0.713	0.776	0.849	0.859	0.860
14	0.715	0.794	0.874	0.875	0.872

III. RESULTS

In the first experiment, five ROC curves of the testing data set are computed as the size of reference library increases (as shown in Fig. 1). Figure 2 demonstrates the trend of five A_z values computed from the corresponding ROC curves. It shows that initially as the size of reference library increases from 630 (library 1) to 1891 ROIs (library 3) the performance of the I-CAD scheme in classifying between true-positive and false-positive mass regions monotonically and significantly improves ($p < 0.01$) with A_z values increasing from 0.715±0.026 to 0.874±0.017. Then, I-CAD performance plateaus. Increasing the size of reference library (adding more reference ROIs) beyond this point does not improve I-CAD performance. There is no statistically significant difference in I-CAD performance between using reference libraries 3, 4, and 5 ($p > 0.9$). Similar trends in I-CAD performance as reference library size increases are found when different numbers of features are used in the KNN algorithm (as shown in Table II).

In the second experiment, we selected four sets of threshold values for true-positive and false-positive mass regions as shown in Table III. The table also summarizes: (1) the number of ROIs (including both true-positive and false-positive) retained in the reference library after removing the poorly effective reference regions, (2) the performance of I-CAD scheme on the same independent testing data set (including A_z values and the corresponding standard deviations), and (3) the computed two-tailed P value between each new ROC curve after applying a pair of thresholds to remove a fraction of poorly effective reference ROIs and the ROC curve generated from the original reference library with a total of 3153 ROIs. The results indicate that as we remove a fraction of poorly effective ROIs from the reference library, the I-CAD performance significantly improves ($p < 0.01$) as shown in Table III. The maximum performance of I-CAD

scheme in this experiment is achieved by setting the two threshold values to be 0.15 for true-positive mass regions and 0.85 for false-positive regions, respectively. Using this pair of thresholds, a total of 5.5% (174 out of 3153) ROIs are identified as poorly effective ROIs (including 85 true-positive regions and 89 false-positive regions). By removing these 174 ROIs from the reference library, 74 false-positive regions and 61 true-positive regions in the testing data set exhibit no change in their detection scores. However, the detection scores of 265 testing ROIs (66% of 400) are changed, which indicates that removing these 174 poorly effective ROIs changes at least one of the 15 similar reference ROIs selected for each of these 265 testing ROIs. Figure 3 shows the histogram of the detection score changes and demonstrates that more false-positive mass regions tend to reduce detection scores, while more true-positive mass regions tend to increase detection scores. Specifically, the average detection scores are 0.634 and 0.267 for 200 true-positive and 200 false-positive testing regions using the original reference library, respectively. By removing the 174 poorly effective reference ROIs, the two average detection scores are changed to 0.684 for true-positive testing regions (7.9% increase) and 0.252 for false-positive testing regions (5.6% decrease). Compared to the original reference library with 3153 ROIs (as shown in the first experiment), using this new library with 2979 reference ROIs significantly improves I-CAD performance ($p < 0.01$) by increasing the A_z value from 0.872±0.017 to 0.914±0.012 (as shown in Fig. 4).

Table IV summarizes and compares I-CAD performance after we applied the same thresholds (0.15 for true-positive masses and 0.85 for false-positive regions) to identify and remove poorly effective ROIs in five reference libraries as shown in Table I. The results show that while removing the fraction of poorly effective ROIs improves I-CAD performance, I-CAD performance still improves as reference li-

TABLE III. I-CAD performance as the change of two threshold values selected to identify and remove poorly effective ROIs from the original reference library. (Note: TP—true-positive, FP—false-positive, and STD—standard deviation.)

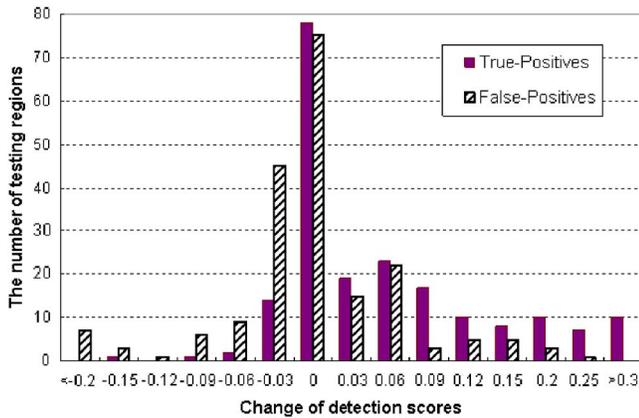| Threshold for TP ROIs	Threshold for FP ROIs	Number of TP ROIs	Number of FP ROIs	Total reference ROIs	A_z value and STD	P value to the original library
N/A	N/A	1592	1561	3153	0.872±0.017	N/A
0.05	0.95	1561	1538	3098	0.901±0.012	0.002
0.15	0.85	1507	1472	2979	0.914±0.012	0.001
0.25	0.75	1439	1363	2802	0.904±0.014	0.006
0.35	0.65	1346	1230	2576	0.887±0.015	0.213

FIG. 3. The histogram of the change of detection scores when applying the I-CAD scheme to the testing data set between using the original reference library and the optimal library by removing 174 poorly effective reference ROIs.

brary size increases. Hence, in order to achieve optimal performance of I-CAD systems that use CBIR schemes, one should first increase the size of the reference library and then identify and remove the poorly effective ROIs.

IV. DISCUSSION

Current CAD schemes for medical images are typically trained using a global based machine learning method (e.g., an artificial neural network) that generates a single function hypothesis (global approximation) that covers the entire instance space and all future testing cases. It is well known that increasing the size and diversity of the training data set gen-
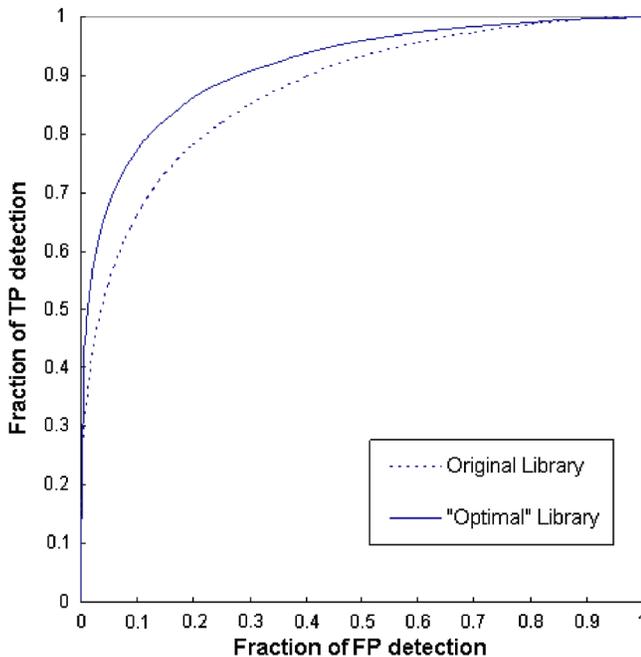

FIG. 4. Two ROC curves of I-CAD scheme when applying to the same testing data set using either the original library involving 3153 reference ROIs or the optimal library after removing 174 poorly effective reference ROIs. The A_z values are 0.872 ± 0.017 and 0.914 ± 0.012, respectively.

erally improves the performance and robustness of these CAD schemes when applying to the independent testing data set.[24] However, I-CAD systems that use CBIR schemes and allow interaction between radiologists and computer schemes use local instance-based machine learning methods and share two common characteristics. First, they are lazy learning methods in that the decision of how to generalize beyond the training data is deferred until a new query (instance) is observed. Second, the new query is classified by analyzing a small set of similar instances while ignoring others that are quite different from the query.[25] Although a CBIR scheme may more effectively use a richer hypothesis space by generating many local functions that are combined to form its implicit global approximation to the target function, the disadvantages of the instance-based learning method may make CBIR scheme more sensitive to the local image noise (poorly effective references). As a result, building effective and concise reference libraries (data reduction) for local instance-based machine learning systems has attracted wide research interests in the field of computational intelligence[26,27] including developing CBIR schemes of mammograms.[17] In this study we assembled a larger image reference library compared to the previously reported studies[9–15] and an independent testing data set. Using this reference library and testing data set, we systematically investigated the relationship between reference library size and I-CAD performance. We also developed a straightforward reference selection method to improve I-CAD performance by identifying and removing a fraction of poorly effective references. I-CAD performance is evaluated based on an independent and relatively large testing data set in this study, which substantially reduces the testing bias and improves the robustness of the testing results.

The results of our first experiment (Figs. 1 and 2) show that as the size of the reference library increases, the I-CAD performance in classification between true-positive and false-positive mass regions significantly improves ($p < 0.01$) with the area under ROC curve increasing monotonically before it reaches the maximum performance level (plateaus). Although the number of features used in the CBIR scheme affects the I-CAD performance level, the trend of I-CAD performance with increasing reference library size remains the same (as shown in Table II), which suggests that such a trend is largely independent of the number of features used in the multifeature based CBIR scheme. This finding indicates that initially adding more reference ROIs increases the signal-to-noise ratio of database and as the signal-to-noise ratio gradually plateaus, arbitrarily increasing the size of the image reference library does not further improve I-CAD performance. Hence, although in the current field of medical imaging the number of available clinical images (or examinations) quickly increases with advances in digital imaging and computing technologies and systems, adding new clinical images (or examinations) into the reference library is an important and difficult issue that affects the performance of CBIR schemes. Researchers have recognized that more so-

TABLE IV. I-CAD performance improvement by removing poorly effective ROIs from five reference libraries with different sizes.

Library	1	2	3	4	5
A_Z using original library	0.715	0.794	0.874	0.875	0.872
The number of deleted ROIs	32	48	96	106	174
A_Z using new library	0.721	0.819	0.890	0.899	0.914
P value of two A_Z values	0.259	0.004	0.058	0.003	<0.001

phisticated reference selection strategies should be investigated and developed to solve this difficult issue.[17]

We investigated in this study a simple strategy to identify the poorly effective ROIs and removed them from the reference library. The proposed hypothesis of that a new incoming ROI should be removed if it performs poorly by comparing it to a group of the similar ROIs in a large and diverse reference library has been successfully validated in our second experiment using the independent testing data set. Due to the use of local instance-based machine learning method, the poorly effective ROIs identified by the CBIR schemes are not the outliers of the global database (reference library). Whether a reference ROI is considered poorly effective is determined by only a limited number of its nearest neighbors (e.g., 15 in this study). Specifically, a poorly effective ROI is dominantly surrounded by the opposite (wrong) class of ROIs (e.g., a true-positive ROI surrounded by a group of false-positive ROIs). It is a regionally misfitted ROI. Unlike the outliers identified in the global database, the poorly effective ROIs identified by the CBIR schemes can widely spread in the different locations of the multidimensional feature space domain (e.g., 14 dimensions in this study). As a result, removing these poorly effective ROIs has much more impact in the testing data set than removing the global outliers in which the most of testing ROIs will not select the outliers as the similar reference ROIs.

We recognize that a poorly effective reference ROI identified by the CBIR scheme can represent either a subtle case (quite different from the most of the typical cases acquired from the routine examinations) or a noisy ROI with distorted image content (e.g., computed features). To avoid or minimize the risk of removing subtle cases in building a concise and effective reference library, it is critical to start with an initially large and diverse image reference library. In other words, including more diverse and subtle cases in the original reference library reduces the probability of the subtle true-positive cases being dominantly surrounded by a group of "typical" K false-positive ROIs (e.g., $K=15$ in this study) in an optimized multidimensional (feature) space domain. In this study, we assembled a large reference library involving 3153 ROIs that were extracted from different image categories with diverse image characteristics (including the masses that are depicted on the images acquired from the "current," "prior," and false-negative examinations, and the false-positive regions that are detected by CAD scheme from positive, recalled, and screening negative examinations[28]). Hence, we reduce the probability of a subtle true-positive mass being dominantly surrounded by K false-positive ROIs

in our reference library. We also recognize that visual subtleness of mass regions is a subjective concept with large inter-reader variability[16] and using combination of a set of computed image features may not accurately correlate to the level of visual subtleness. Because the computed image features are affected by the image noise depicted on the original digitized mammograms and mass region segmentation error, our data analysis shows the large image feature variation of the poorly effective ROIs removed from the reference library. Although we conjecture that the identified poorly effective ROIs are mainly noisy and distorted reference samples due to variety of errors in image acquisition techniques and image processing methods (including mass segmentation error), we cannot rule out the possibility of that some of the removed ROIs are truly subtle cases. How to effectively remove noisy (distorted) reference ROIs while avoiding or minimizing the risk of removing subtle cases is an important and unsolved issue in building the concise and effective reference libraries used in CBIR schemes. Further research is needed to investigate this issue and find the optimal solution. Despite this limitation of our preliminary study, the testing results are encouraging. In particular, by repeating experiment one using the CBIR scheme with different number of features (as shown in Table II) and experiment two by identifying and removing poorly effective ROIs from five reference libraries with different sizes (as shown in Table IV), we demonstrate the generalizability of (1) the trend between the I-CAD performance and the increase of reference library size and (2) the effectiveness of the proposed reference selection strategy (removing poorly effective reference ROIs) to improve I-CAD performance.

Two typical types of CBIR schemes have been developed and extensively tested in I-CAD schemes of mammograms. One uses multi-image features and distance-weighted KNN algorithm and the other uses an information-theoretic (IT) based template matching scheme to search for the similar reference images (or ROIs). Both approaches have advantages and disadvantages. First, lesion (including breast mass) segmentation is always a very difficult task and its result substantially affects the accuracy of computed image features. Hence, due to the lack of a reliable (or robust) algorithm for segmenting masses surrounded and overlapped by complex (e.g., heterogeneously dense) breast tissues using two-dimensional (projective) mammograms, many of previous studies used the semiautomated method with manual correction to segment identified mass regions with fuzzy boundary in an attempt to improve accuracy of computed image features representing image content.[14,19] Due to large inter-

reader variability in segmentation of lesion boundary contours,[29] semiautomated segmentation can substantially reduce but not eliminate segmentation errors. Using IT-based template matching schemes can avoid the difficulty and error of lesion segmentation. However, the performance of these schemes is affected by the selection of ROI size and the number of histogram bins (i.e., using mutual information based template matching).[30] When ROIs with fixed size [i.e., 512×512 pixels (Ref. 13)] are used, the similarity of ROIs depicting small masses may also largely depend on the similarity of the surrounding normal breast tissues rather than mass content. Second, computational efficiency is another important factor that determines the clinical utility of the I-CAD systems in assisting radiologists to interpret and diagnose medical images. Because all image features that represent reference ROIs are precomputed (off-line), the CBIR scheme using a multifeature based KNN algorithm can be executed to search for the similar reference ROIs, compute the detection score, and display the results in I-CAD workstation in real time.[16] IT-based CBIR schemes are more computationally expensive because each reference image (ROI) must be processed pixel-by-pixel on-line, which makes interactive use by radiologists difficult. Therefore, due to the limitations of both types of CBIR schemes, more investigation and development is needed to more accurately segment lesions (or determine ROI size), extract effective image features (or content), increase computational efficiency, and assemble an optimal reference library.

In addition, we believe that the size and diversity are the two most important factors to determine the effectiveness of reference library (database) used in the CBIR schemes. Due to the large variability of breast masses depicted on clinical (screening and diagnostic) mammograms, without a sufficiently large size, one cannot build a diverse database. However, the large size does not guarantee the diversity of the database. How to objectively and reliably assess the diversity of the reference libraries is another difficult and unsolved issue in the optimization of CBIR schemes. In our future study we will focus to develop and compare different methods and criteria to assess database diversity. Then, we will investigate how the diversity level of the database affects the size of the optimal reference libraries used in CBIR schemes.

V. CONCLUSION

The quality or effectiveness of the image reference library plays a critical role in developing I-CAD systems using CBIR schemes. Assembling and evaluating an optimal reference library remains a difficult technical challenge in this area. In this study, we conducted two experiments to investigate (1) the relationship between the size of reference library and I-CAD performance in classification between true-positive and false-positive breast masses and (2) the possibility of improving I-CAD performance by identifying and removing a small number of poorly effective reference ROIs using a new proposed reference selection strategy. The experimental results indicate that in order to build a highly and robustly performed CBIR scheme, one first needs to as-

semble a large and diverse reference library. Once the size of the reference library reaches a level at which the performance of the CBIR scheme plateaus, continuing to add arbitrary references into the library will typically not help to further improve scheme performance. In order to improve performance of CBIR schemes, it is important to identify and remove poorly effective reference regions from the library and to examine the effectiveness of each new incoming image sample before depositing it to the reference library.

[a)]Author to whom correspondence should be addressed. Telephone: (412)-641-2568; Fax: (412)-641-2582. Electronic mail: zhengb@upmc.edu

[1]H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—Clinical benefits and future directions," Int. J. Med. Inform. **73**, 1–23 (2004).

[2]H. Muller *et al.*, "Benefits of content-based visual data access in radiology," Radiographics **25**, 849–858 (2005).

[3]R. M. Nishikawa, "Current status and future directions of computer-aided diagnosis in mammography," Comput. Med. Imaging Graph. **31**, 224–235 (2007).

[4]T. M. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," Radiology **220**, 781–786 (2001).

[5]L. A. Khoo, P. Taylor, and R. M. Given-Wilson, "Computer-aided detection in the United Kingdom National Breast Screening Programme: Prospective study," Radiology **237**, 444–449 (2005).

[6]J. M. Ko, M. J. Nicholas, J. B. Mendel, and P. J. Slanetz, "Prospective assessment of computer-aided detection in interpretation of screening mammograms," AJR, Am. J. Roentgenol. **187**, 1483–1491 (2006).

[7]R. M. Nishikawa and M. Kallergi, "Computer-aided detection in its present form is not an effective aid for screening mammography," Med. Phys. **33**, 811–814 (2006).

[8]J. J. Fenton *et al.*, "Influence of computer-aided detection on performance of screening mammography," N. Engl. J. Med. **356**, 1399–1409 (2007).

[9]M. L. Giger *et al.*, "Intelligent CAD workstation for breast imaging using similarity to known lesions and multiple visual prompt aides," Proc. SPIE **4684**, 768–773 (2002).

[10]I. El-Naqa *et al.*, "A similarity learning approach to content-based image retrieval: Application to digital mammography," IEEE Trans. Med. Imaging **23**, 1233–1244 (2004).

[11]C. Wei, C. Li, and R. Wilson, "A general framework for content-based medical image retrieval with its application to mammograms," Proc. SPIE **5748**, 134–143 (2005).

[12]H. Alto, R. M. Rangayyan, and J. E. Desautels, "Content-based retrieval and analysis of mammographic masses," J. Electron. Imaging **14**, 023016-1–17 (2005).

[13]G. D. Tourassi *et al.*, "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," Med. Phys. **34**, 140–150 (2007).

[14]Y. Tao, S. B. Lo, M. T. Freedman, and J. Xuan, "A preliminary study of content-based mammographic masses retrieval," Proc. SPIE **6514**, 65141Z-1–12 (2007).

[15]C. Muramatsu *et al.*, "Determination of subjective similarity for pairs of masses and pairs of clustered microcalcifications on mammograms: Comparison of similarity ranking scores and absolute similarity ratings," Med. Phys. **34**, 2890–2895 (2007).

[16]B. Zheng *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).

[17]G. D. Tourassi, B. Harrawood, S. Singh, and J. Y. Lo, "Information-theoretic CAD system in mammography: Entropy-based indexing for

computational efficiency and robust performance," Med. Phys. **34**, 3193–3204 (2007).

[18] B. Zheng *et al.*, "Computer-aided detection schemes: The effect of limiting the number of cued regions in each case," AJR, Am. J. Roentgenol. **182**, 579–583 (2004).

[19] B. Zheng *et al.*, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," Med. Phys. **33**, 111–117 (2006).

[20] B. Zheng, Y. H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single image segmentation and a multilayer topographic feature analysis," Acad. Radiol. **2**, 959–966 (1995).

[21] L. Yang *et al.*, "Learning distance metrics for interactive search-assisted diagnosis of mammograms," Proc. SPIE **6514**, 65141H-1–12 (2007).

[22] C. E. Metz, ROCFIT 0.9B Beta version, http://www.radiology.uchicago.edu/krl/, University of Chicago, Chicago, IL, 1998.

[23] Q. Li and K. Doi, "Reduction of bias and variance for evaluation of computer-aided diagnostic schemes," Med. Phys. **33**, 868–875 (2006).

[24] B. Zheng, Y. H. Chang, W. F. Good, and D. Gur, "Adequacy testing of training set sample sizes in development of a computer-assisted diagnosis scheme," Acad. Radiol. **4**, 497–502 (1997).

[25] T. M. Mitchell, *Machine Learning* (WCB McGraw-Hill, Boston, MA, 1997).

[26] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 873–885 (1989).

[27] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artif. Intell. **97**, 245–271 (1997).

[28] D. Gur *et al.*, "CAD performance on sequentially ascertained mammographic examinations of masses: An assessment," Radiology **233**, 418–423 (2004).

[29] A. P. Reeves *et al.*, "The lung image database consortium (LIDC): Pulmonary nodule measurements, the variation, and the difference between different size metrics," Proc. SPIE **6514**, 65140J-1–8 (2007).

[30] P. Filev *et al.*, "Comparison of similarity measures for the task of template matching of masses on serial mammograms," Med. Phys. **32**, 515–529 (2005).