

Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms

Georgia D. Tourassi^{a)} and Brian Harrawood

Digital Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, North Carolina 27705

Swatee Singh, Joseph Y. Lo, and Carey E. Floyd

Digital Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, North Carolina 27705 and Department of Biomedical Engineering, Duke University, Durham, North Carolina 27710

(Received 14 June 2006; revised 3 November 2006; accepted for publication 6 November 2006; published 18 December 2006)

The purpose of this study was to evaluate image similarity measures employed in an information-theoretic computer-assisted detection (IT-CAD) scheme. The scheme was developed for content-based retrieval and detection of masses in screening mammograms. The study is aimed toward an interactive clinical paradigm where physicians query the proposed IT-CAD scheme on mammographic locations that are either visually suspicious or indicated as suspicious by other cueing CAD systems. The IT-CAD scheme provides an evidence-based, second opinion for query mammographic locations using a knowledge database of mass and normal cases. In this study, eight entropy-based similarity measures were compared with respect to retrieval precision and detection accuracy using a database of 1820 mammographic regions of interest. The IT-CAD scheme was then validated on a separate database for false positive reduction of progressively more challenging visual cues generated by an existing, in-house mass detection system. The study showed that the image similarity measures fall into one of two categories; one category is better suited to the retrieval of semantically similar cases while the second is more effective with knowledge-based decisions regarding the presence of a true mass in the query location. In addition, the IT-CAD scheme yielded a substantial reduction in false-positive detections while maintaining high detection rate for malignant masses. © 2007 American Association of Physicists in Medicine.

[DOI: [10.1118/1.2401667](https://doi.org/10.1118/1.2401667)]

I. INTRODUCTION

There is conflicting evidence regarding the clinical impact of computer-assisted detection (CAD) systems for the diagnostic interpretation of screening mammograms. For the most part, retrospective studies suggest that CAD technology has a positive impact on early breast cancer detection (e.g., Refs. 1–5). There are, however, several retrospective^{6–8} and prospective^{9–13} studies that produced contradictory conclusions. Although it is recognized that more prospective studies are needed on the topic, it is well known that radiologists often dismiss correct CAD cues. The radiologists' reluctance to trust CAD is mainly attributed to the higher than desired false positive rate.¹¹ The above observations are particularly true for the detection of masses, a far more challenging task than the detection of calcifications.

While the true clinical benefit of CAD is still debated,¹⁴ CAD research continues in an effort to improve diagnostic performance and clinical integration.¹⁵ For example, the currently used “black-box” CAD paradigm is rather limited. A CAD system that is more interactive and capable of justifying the visual cues it provides may help radiologists' cognitive process more effectively. Moreover, as clinical image libraries grow rapidly in Radiology, contemporary CAD systems should be able to capitalize on accumulating image data without requiring painstaking retraining or recalibration.

Content-based image retrieval (CBIR) could facilitate the development of a new generation of interactive CAD technology that takes advantage of the vast amounts of digital image data generated in clinical practice. The main objective of CBIR research is to develop a user-friendly framework that allows users to interact with digital image libraries effectively.¹⁶ CBIR has been identified as an important research direction in Radiology to facilitate clinical decision support for medical image interpretation.^{17,18}

Shifting the CAD paradigm to incorporate image retrieval capabilities is a challenging proposition. The primary task of CBIR in the clinical arena is to help radiologists retrieve images with similar visual content. Medical image retrieval has traditionally been based on text describing the patient clinical data and medical condition depicted in the patient's imaging studies. These textual descriptors are used as keywords for searching the medical image library. Several researchers have recognized the need for more sophisticated image retrieval methods that capture the visual content of images more effectively than textual descriptors. Consequently, CBIR has evolved toward feature-based similarity assessment. Images are compared and retrieved based on low-level image features that describe the color, shape, texture, and spatial arrangement of important objects (i.e., organs, tumors, etc.) identified in the medical images. Nevertheless, low-level image features are often ineffective in

CBIR of single-modality images due to the subtle differences that exist among same-domain images. This inefficiency is known as the “semantic gap” between image features and the visual and diagnostic content of the images as perceived by the radiologists.¹⁷ Therefore, the challenge in creating clinically effective CBIR-based CAD systems is to develop algorithms that retrieve semantically and perceptually similar images to provide evidence-based decision support.

Working toward this goal, we have previously presented a CBIR-based CAD system for the detection and diagnosis of masses in screening mammograms.^{19,20} In contrast to feature-based CBIR algorithms in mammography,^{21–27} our system relies on information theoretic principles to assess image similarity. Specifically, the system uses the popular concept of mutual information (MI) to measure the similarity between a query image and those stored in the knowledge database. MI-based similarity assessment relies completely on the statistical properties of the image histograms eliminating the image preprocessing, segmentation, and feature extraction steps. Furthermore, information theoretic similarity measures have the advantage of making no assumptions on the underlying image distributions. Our CAD system was evaluated initially as a knowledge-based system for the discrimination of masses from normal breast parenchyma¹⁹ and for the diagnostic characterization of masses using relevance feedback techniques.²⁰

Since similarity assessment is the most important component in CBIR,^{28,29} the purpose of this study was to explore several entropy-based similarity measures for region-based analysis of mammograms. Specifically, we present a comparative study using the information-theoretic computer-assisted detection (IT-CAD) scheme for three clinically oriented tasks. First, an experiment was performed to determine which similarity measure helps the IT-CAD scheme retrieve semantically relevant mammographic regions with the highest precision. A second experiment was performed to determine which measure helps the IT-CAD scheme discriminate between mass and normal mammographic regions with the highest accuracy. Finally, a third experiment was performed to validate the conclusions of Experiments 1 and 2 using IT-CAD for evidence-based, false positive reduction of progressively more challenging visual cues produced by an existing second-reader CAD system.

II. MATERIALS AND METHODS

A. Information-theoretic similarity measures

Information-theoretic (dis)similarity measures are based on the concept of entropy.³⁰ The most commonly used entropy definition is the Shannon entropy (H):

$$H = - \sum_x p(x) \log_2[p(x)], \quad (1)$$

where $p(x)$ is the probability that an image pixel will have the intensity value x . The image probability $p(x)$ is typically estimated from the image histogram, commonly using the convention $0 \log 0 = 0$. Entropy is considered a measure of the uncertainty or complexity in an image. The image com-

plexity (or uncertainty) is captured by the dispersion of the probability distribution of the image intensity levels. Images with uniform pixel intensity distributions have high dispersion and therefore higher entropy. In contrast, images with intensity distributions that depict a few large peaks have lower dispersion and thus lower entropy.

Generally, information-theoretic similarity measures compare the histograms of two images X and Y . The comparison may focus only on corresponding histogram bins (i.e., bin-by-bin measures) or it may incorporate information for non-corresponding bins (i.e., cross-bin measures). This study investigates eight information-theoretic (IT) (dis)similarity measures that have been successfully applied in other areas of medical imaging such as image registration, segmentation, and feature-based image retrieval. Four of them are cross-bin measures: (i) joint entropy, (ii) conditional entropy, (iii) mutual information, and (iv) normalized mutual information. The remaining four IT measures are typical examples of bin-by-bin measures: (i) average Kullback-Leibler divergence, (ii) maximum Kullback-Leibler divergence, (iii) Jensen divergence and, (iv) arithmetic-geometric mean divergence. The following is a brief description of each measure.

1. Joint entropy

Joint entropy (JOINT_H) is the entropy of the joint histogram of two images X and Y .

$$JOINT_H = H(X, Y) = - \sum_x \sum_y p_{XY}(x, y) \log[p_{XY}(x, y)]. \quad (2)$$

If two images are completely unrelated, their joint entropy is equal to the sum of their individual entropies. On the other hand, the more similar two images are, the lower their joint entropy is compared to the sum of the individual entropies. Consequently, the joint entropy is a distance measure rather than a similarity measure. Two images with lower joint entropy are considered more similar (i.e., more relevant) than two images with higher joint entropy.

2. Conditional entropy

The conditional entropy $H(X|Y)$ of two images X and Y measures how much entropy (or uncertainty) is remaining regarding image X (i.e., the query image) when we have learned the truth regarding image Y (i.e., an image in the knowledge database). Similarly to joint entropy, conditional entropy is also a dissimilarity measure. Therefore, if two images are relevant, then the conditional entropy (or uncertainty) of the query image given the known image should be low. However, in contrast to joint entropy, conditional entropy is not symmetric. In other words, $H(X|Y) \neq H(Y|X)$. The conditional (COND_H) and joint entropy of two images X and Y are related as follows:

$$COND_H = H(X|Y) = H(X, Y) - H(Y) \quad (3)$$

3. Mutual information

Mutual information (MI) is the most popular IT similarity measure, particularly for image registration.^{31–33} MI is similar to joint entropy but it also takes into account the individual image entropies.

$$\begin{aligned} MI(X, Y) &= H(X) + H(Y) - H(X, Y) \\ &= \sum_x \sum_y P_{XY}(x, y) \log_2 \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)}. \end{aligned} \quad (4)$$

MI is a measure of general statistical dependence (i.e., shared information) between two images. It measures the amount of uncertainty reduction about one image given the information we have about the other image. MI is a true similarity measure. The more similar X and Y are, the higher their MI. Furthermore, MI is considered a generalized extension of the correlation coefficient because it does not make linear assumptions regarding the relationship between the two images' pixel values.³⁴

4. Normalized mutual information

Normalized mutual information (NMI) is a normalized version of MI ensuring that the similarity measure is bounded between 0 and 1. Previous studies in image registration have shown that NMI is often more successful and robust than MI (Ref. 31).

$$NMI(X, Y) = \frac{H(X) + H(Y)}{H(X, Y)}. \quad (5)$$

5. Relative entropy

Relative entropy or Kullback-Leibler (KL) divergence is a distance measure between two probability distributions $p(x)$ and $q(x)$. In the scope of this study, $p(x)$ and $q(x)$ are the probability distributions of the stored image $p(x)$ and the query image $q(x)$, respectively. The relative entropy is defined as follows:

$$D(q \parallel p) = \sum_x q(x) \log[q(x)/p(x)]. \quad (6)$$

Relative entropy is typically used in coding theory and it measures how inefficient on average it would be to use the histogram of one image to code another. Generally, the higher the relative entropy is, the more dissimilar the two images are. Similarly to conditional entropy, KL divergence is not a true distance measure because it is not symmetric (i.e., $D(q \parallel p) \neq D(p \parallel q)$). Consequently, different transformations have been utilized in CBIR to provide a symmetric KL divergence measure (SKL).³⁵ In this study, we have explored two such transformations: (i) the average KL divergence

$$SKL_{-1} = \frac{D(q \parallel p) + D(p \parallel q)}{2} \quad (7)$$

and (ii) the maximum KL divergence

$$SKL_{-2} = \max[D(q \parallel p), D(p \parallel q)]. \quad (8)$$

SKL is a non-negative distance metric that is equal to 0 when the two probability distributions are identical.

6. Jensen divergence

Some studies have indicated that KL divergence $D(p \parallel q)$ is not numerically stable and is often sensitive to histogram binning.³⁶ Consequently, another divergence measure has been proposed as a more stable alternative. The Jensen divergence (JD) is an empirical modification of the KL divergence that is symmetric and more robust with respect to noise and histogram binning³⁶

$$\begin{aligned} JD(p, q) &= \sum_x \left(q(x) \log \frac{2q(x)}{p(x) + q(x)} \right. \\ &\quad \left. + p(x) \log \frac{2p(x)}{p(x) + q(x)} \right). \end{aligned} \quad (9)$$

The Jensen divergence has values bounded between 0 and 2.

7. Arithmetic-geometric mean divergence

Finally, the last similarity measure explored was the arithmetic-geometric mean (AGM) divergence. This measure is essentially the KL divergence between the arithmetic and geometric means of the two image distributions $p(x)$ and $q(x)$

$$AGM(p, q) = \sum_x \frac{p(x) + q(x)}{2} \log \frac{p(x) + q(x)}{2\sqrt{p(x)q(x)}}. \quad (10)$$

All above IT measures require estimation of the marginal probability distribution of the individual images. In addition, some measures (i.e., JOINT_H, COND_H, MI, NMI) require estimation of the joint probability distribution of the two images as well. Consistent with our earlier study¹⁹ and for reasons of computational efficiency, we applied the histogram approach³³ to approximate the marginal and joint probability distributions functions. The number of histogram bins for histogram approximation was selected empirically. We varied the number of histogram bins (i.e., 4, 8, 16, 32, 64, 128) and repeated the experiments with respect to all similarity measures. As expected, the number of histogram bins affected the observed results. For example, using only four bins produced consistently inferior results across all similarity measures. The differences among the results observed for the remaining values of the histogram bin parameter were not statistically significant. Overall, 64 bins were sufficient for histogram approximation across all similarity measures and clinical tasks. For each ROI, the mean μ and standard deviation σ of the ROI pixel values were calculated. Then, the interval $[\mu - 2\sigma, \mu + 2\sigma]$ was divided into 64 equal-sized bins. Pixel values falling outside the predetermined $[\mu - 2\sigma, \mu + 2\sigma]$ interval were assigned to the outermost bins when calculating the histograms. The above rules were followed consistently for all images, similarity measures, and experiments.

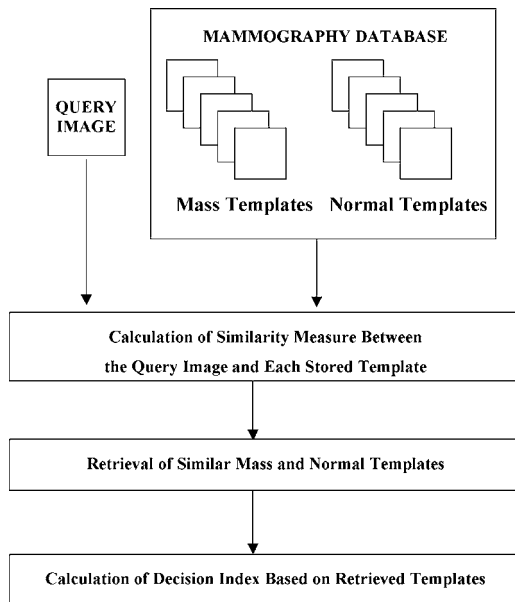


FIG. 1. Schematic representation of the IT-CAD content-based retrieval and detection scheme for mammographic masses.

B. Overview of the information-theoretic CAD system

Figure 1 shows a schematic representation of the image retrieval scheme with the proposed information theoretic framework for mass detection. The scheme is designed to provide region-based evaluation of mammograms for a targeted, evidence-based analysis of suspicious mammographic locations.

Initially, a query mammographic location is presented to the IT-CAD system. The system extracts a fixed size mammographic region around the specific location. The query region of interest (ROI) is compared to a knowledge database of ROIs with known ground truth. Similar cases are retrieved from the knowledge database. A decision is formulated regarding the query region using the retrieved similar cases.

There are two critical components in the IT-CAD scheme: (i) the similarity measure, and (ii) the knowledge database. Since the clinical focus of the IT-CAD scheme is mass detection, it is reasonable to expect that the knowledge database should contain a rich collection of mammographic ROIs that depict biopsy-proven masses. Although the above requirement is critical, the knowledge database also includes a diverse set of ROIs that depict normal breast parenchyma. Because the similarity measure is calculated using the full ROI, it is possible that two ROIs may result in high similarity mainly due to parenchymal background similarities rather than the potential abnormalities they contain. Consequently, the information theoretic CAD approach decides based on both similar mass and normal cases that are stored in its knowledge database. Specifically, the IT-CAD decision index $D(Q)$ is calculated as follows:

$$D(Q) = \frac{1}{k} \sum_{j=1}^k SM(Q, M_j) - \frac{1}{k} \sum_{j=1}^k SM(Q, N_j), \quad (11)$$

where Q is the query mammographic region, SM stands for similarity measure, and M_j and N_j are known mass and normal cases that are retrieved from the knowledge database as most similar to the query. Note that if the query region depicts a mass, it is expected that the calculated decision index should be higher than if it contains normal parenchyma. The second term in Eq. (11) is a correction term so that high values of $D(Q)$ are less likely to be the result of matching backgrounds than matching potential abnormalities.

Although our previous studies have shown promising results using mutual information as the similarity measure (SM), this study explores several other information-theoretic (dis)similarity measures that share the same featureless simplicity and computational efficiency with MI. Note that in Eq. (11), SM denotes a similarity measure. The proposed dissimilarity measures (i.e., joint entropy, conditional entropy, KL divergence, Jensen divergence, and geometric/arithmetic mean) can be easily converted into similarity measures by taking their negative or inverse value. For this study, we applied the negative transformation.

C. Data collection and study design

The study was based on 512×512 pixel ROIs extracted from mammograms for the Digital Database of Screening Mammography (DDSM).³⁷ The mammograms are 12 bit images digitized using the Lumisys scanner at $50 \mu\text{m}$ per pixel. No image preprocessing (i.e., segmentation, filtering, normalization, etc.) was performed on the mammograms or the extracted ROIs.

We created two different ROI databases based on the DDSM/Lumisys mammograms. Database 1 contained 1820 ROIs. Of those, 901 ROIs depicted a biopsy-proven mass (489 malignant and 412 benign). The ROIs were centered around the physician's annotation provided in the DDSM truth files. The remaining ROIs were extracted from 62 normal mammograms (two ROIs per breast, per view) for a total of $8 \times 62 = 496$ normal ROIs. The location of the normal ROIs was selected randomly within the breast. There was no overlap between the ROIs extracted from the same image. To keep the database evenly balanced between normal and abnormal ROIs, an additional 424 ROIs were extracted from abnormal DDSM/Lumisys cases, but only from breasts that did not contain any physician annotations in either mammographic view. The selection of these cases was random. Therefore, the final database contained 919 ROIs that were deemed normal.

Database 2 contained ROIs extracted from 100 DDSM cases completely different from those used to create Database 1. This second database was selected to represent a balanced mix of abnormal and normal cases from all available DDSM/Lumisys volumes. Note that the DDSM volumes correspond to patient data acquired at different geographic locations. By creating a balanced mix of cases we tried to minimize potential discrepancies due to patient popu-

lation differences. Furthermore, within each volume an equal number of cases were selected for each mammographic density. Of the 100 DDSM cases in Database 2, 40 cases contained malignant masses, 40 cases contained benign masses, and the remaining 20 cases were considered normal. In DDSM a screening mammogram is considered normal if it does not require any further “follow-up,” it does not contain any annotated abnormalities, and the patient has a normal screening exam at least four years later.

Database 2 was processed using a previously presented, in-house CAD system for mass detection.^{38,39} The system was used to locate suspicious locations within the images. The CAD system is a multi-stage algorithm consisting of a typical sequence of steps: (i) image filtration using a difference of Gaussians filter,^{40,41} (ii) initial localization of suspicious regions detected at high sensitivity using a progressive gray level thresholding procedure, (iii) feature extraction and selection, and (iv) feature-based classification using Fisher’s linear discriminant for false positive reduction of the initial suspicious regions. The prescreening, in-house CAD system was initially trained and optimized on a separate set of DDSM cases, completely different from Database 2. After training and optimization, the system was applied “as is” on Database 2.

Specifically, the in-house, mass detection system was applied on the craniocaudal (CC) views of the 80 cases in Database 2 that contained the annotated masses. For the 20 normal cases in Database 2, only one, randomly selected CC view (left or right breast) was analyzed. Therefore, 100 independent images were analyzed. The automated screening process resulted in 399 false positive (FP) detections (approximately 4 FPs/image). In addition, depending on the definition of true positive detection,⁴² the system also detected 84%–92% of the true masses. However, because our main focus is on reducing further the false positive detections, we combined the 399 FPs with all true masses annotated in the 100 images anticipating future sensitivity improvement of our prescreening algorithm. In total, there were 483 mammographic regions in Database 2; 44 depicting a malignant mass, 40 depicting a benign mass, and 399 depicting suspicious looking yet normal breast parenchyma.

Database 1 was used in a leave-one-out manner to assess how the various image similarity measures impact the retrieval precision (Experiment 1) and diagnostic accuracy (Experiment 2) of our IT-CAD scheme. The leave-one-out sampling scheme was implemented on a per case basis as follows. Each ROI in Database 1 was excluded once to serve as the query. Of the remaining 1819 ROIs, the ones extracted from DDSM cases different than the query’s served as the knowledge database of the IT-CAD scheme. The same process was repeated until each ROI served as a query.

Experiment 3 aimed to validate the conclusions drawn from experiments 1 and 2 for the clinical task of reducing the false positive detections of prescreening CAD systems. Both Databases 1 and 2 were used for this third experiment. Specifically, the ROIs in Database 2 served as queries for testing the IT-CAD system while Database 1 served as the knowledge database.

D. Performance evaluation

Two different performance indices were employed in this study depending on the operating mode of the IT-CAD system (retrieval engine vs. detection aid). When the system was tested as a retrieval engine, its retrieval capabilities were assessed using precision as the selected performance index. Given a query image, precision (P) is the number of relevant retrieved images (R) divided by the total number of retrieved images (K)

$$\text{Precision}(P) = \frac{\text{Number of relevant retrieved images } (R)}{\text{Total number of retrieved images } (K)} \quad (12)$$

Retrieval precision in CBIR is analogous to positive predictive value in decision analysis. There are two ways to define relevance in image retrieval; visual or semantic. For this application, we focus on semantic relevance. A retrieved image is considered to be relevant if it belongs to the same class (mass or nonmass) as the query image. Since retrieval precision is dependent on the query, a CBIR system’s precision is typically reported averaged across all queries. According to Eq. (12), retrieval precision is also dependent on the number of retrieved images (K) and it is typically plotted as a function of K . In this study, we focus only on the top 1, 5, and 10 retrievals and evaluate the eight similarity measures with respect to these top retrieved cases. We limited the precision analysis to $K \leq 10$ for practical reasons. In an interactive CAD system, it is impractical to present radiologists with more than the top ten most similar cases for visual evaluation.

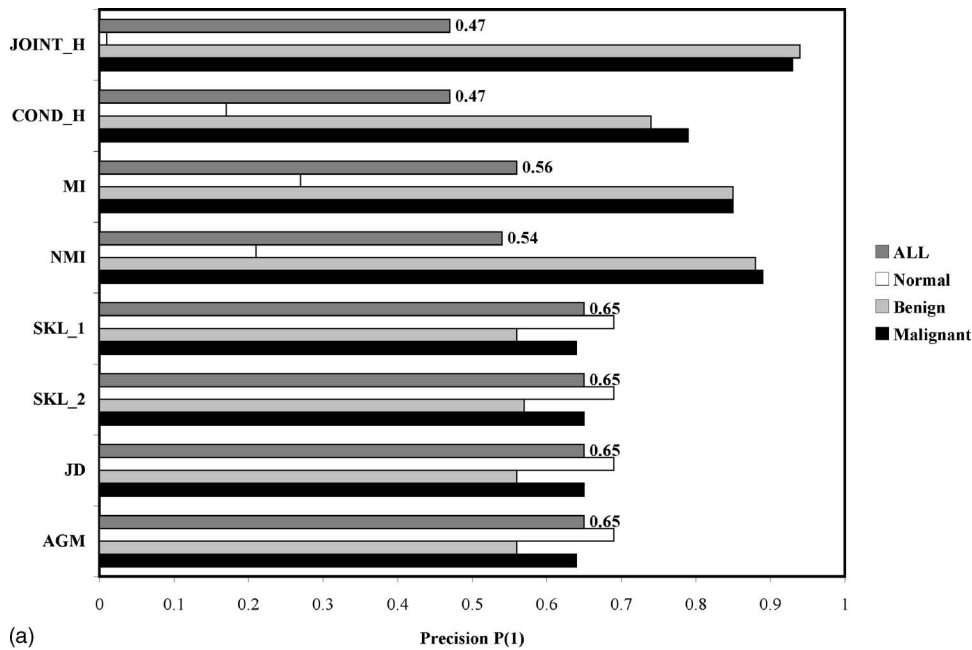
Receiver operating characteristic (ROC) analysis⁴³ was employed to assess the performance of the IT-CAD system as a mass detection aid. The decision index calculated based on Eq. (11) was used as the decision variable for ROC analysis. Since the decision index is dependent on the number of closest mass and normal retrievals (k), ROC analysis was performed for a wide range of k values. The ROC analysis was performed using the ROCKIT software developed by Charles Metz at the University of Chicago.

III. RESULTS

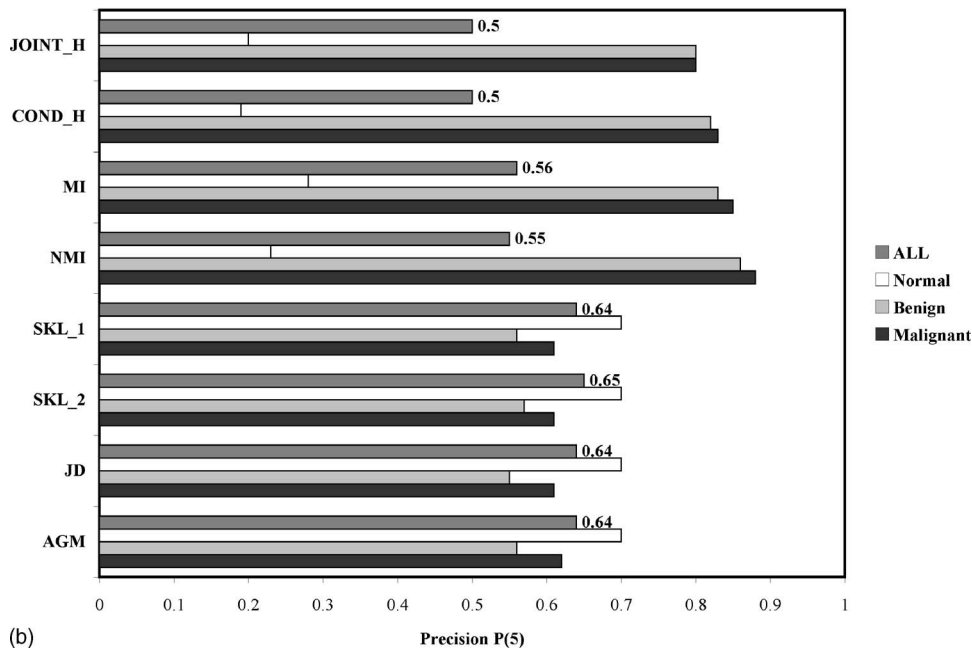
A. Experiment 1: Retrieval precision

The average retrieval precision achieved by each similarity measure at the top $K=1, 5,$ and 10 retrievals was calculated for all queries and separately for each subgroup of queries (i.e., malignant masses, benign masses, and normals). Subgroup analysis was performed to identify possible discrepancies depending on the true class of each query ROI. Overall there were subtle changes as the number of retrievals increased from $K=1$ to $K=10$. Thus, Fig. 2 shows results for the top $K=1$ and $K=5$ retrievals only [Figs. 2(a) and 2(b), respectively].

Figure 2 shows that the overall retrieval precision $P(K)$ achieved by the eight similarity measures appears to be within the range of 47%–65%. Bin-by-bin measures demonstrated overall higher average retrieval precision compared to



(a)



(b)

FIG. 2. Average retrieval precision $P(K)$ for the (a) top $K=1$ and (b) $K=5$ retrievals for all similarity measures. Precision is shown overall and for each subgroup of query cases separately. (JOINT_H: joint entropy, COND_H: conditional entropy, MI: mutual information, NMI: normalized mutual information, SKL: symmetric Kullback-Leibler divergence, SKL_MAX: maximum Kullback-Leibler divergence, JD: Jensen divergence, AGM: arithmetic-geometric mean divergence).

the cross-bin measures at all three retrieval levels. In addition, there were dramatic differences depending on the type of query. Four similarity measures (i.e., joint entropy, conditional entropy, mutual information, and normalized mutual information) achieved significantly higher precision for mass queries rather than normal queries. In contrast, average precision was far more robust between mass and normal queries for the remaining similarity measures (i.e., symmetric Kullback-Leibler divergence, maximum Kullback-Leibler divergence, Jensen divergence, and arithmetic-geometric mean divergence). However, the average retrieval precision was consistently higher for normal queries than masses for the second group of similarity measures. It is notable that the average retrieval precision for malignant masses was consistently

higher than that for benign masses for all similarity measures (with the only exception for the joint entropy measure at the top $K=1$ retrieval).

Note that since the normal and mass ROIs are almost evenly balanced in Database 1, there is a 50% chance to randomly retrieve a mass or normal template from the knowledge database. The Wilcoxon signed rank test was performed to determine if the average precision was significantly higher than the expected $\sim 50\%$ precision value due to the inherent prevalence of each subgroup (i.e., 49.5% for mass and 51.5% for normal ROIs) in the database. For all K values, all similarity measures, and all subgroups of query cases the observed precision was statistically significantly different (p value < 0.0001) than the expected $\sim 50\%$ aver-

TABLE I. ROC area index A_z (± 0.01) achieved by the IT-CAD scheme depending on the similarity measure and the number of the best-matched k mass and normal templates considered in decision making.

k	JOINT_H	COND_H	MI	NMI	SKL_1	SKL_2	AGM	JD
1	0.47	0.48	0.71	0.71	0.68	0.70	0.67	0.70
5	0.48	0.50	0.78	0.79	0.73	0.75	0.74	0.75
10	0.51	0.51	0.81	0.81	0.76	0.77	0.76	0.76
50	0.59	0.57	0.87	0.85	0.77	0.77	0.77	0.77
100	0.67	0.70	0.87	0.85	0.76	0.76	0.75	0.76
300	0.85	0.85	0.86	0.86	0.73	0.73	0.73	0.73
500	0.86	0.86	0.86	0.87	0.63	0.64	0.62	0.65
700	0.87	0.87	0.87	0.87	0.60	0.61	0.60	0.62
ALL	0.86	0.86	0.87	0.87	0.59	0.59	0.58	0.63

age precision if retrieval were purely random. This result was consistent for subgroups and similarity measures where the achieved precision was significantly inferior to that expected with random retrieval (e.g., 20% average precision $P(5)$ for normal ROIs using joint entropy as the similarity measure).

The signed rank test with Bonferroni correction for multiple comparisons was also performed to test for significant differences in average retrieval precision among the different similarity measures. The analysis was performed for each K value ($K=1, 5, 10$) and each query subgroup (malignant, benign, normal) separately at the 95% confidence level. The consistent trend among the results was that the four dissimilarity measures SKL_1, SKL_2, JD, AGM provide very similar average retrieval precision for all query groups.

On the other hand, the remaining four similarity measures (JOINT_H, COND_H, MI, NMI) provide significantly different precision performance compared to the first group across all subgroups and retrieval levels ($K=1, 5, 10$). Indeed, non-parametric correlation analysis confirmed that (SKL_1, SKL_2, JD, AGM) and (JOINT_H, COND_H, MI, NMI) represent two distinct groups of measures. The similarity measures of the first group resulted in highly correlated precision performance for all query groups ($0.87 \leq \rho \leq 0.98$). However, the similarity measures of the second group resulted in significantly less correlated precision performance ($-0.14 \leq \rho \leq 0.62$) with the exception of COND_H and MI ($0.65 \leq \rho \leq 0.90$ depending on the query group and number of top retrievals). Surprisingly, the mutual information and normalized mutual information measures resulted in lower correlation ($0.56 \leq \rho \leq 0.84$ depending on the query group and number of top retrievals). It is noted that the differences in precision between MI, NMI, COND_H, and JOINT_H were often significant for both mass and normal queries at the various retrieval levels. The above statistical analysis was performed using the JMP Statistical Software Version 5.1 available from SAS, Cary, NC.

B. Experiment 2: Detection accuracy

The (dis)similarity measures were subsequently used in the IT-CAD system for the discrimination of mass from normal ROIs according to the decision variable described in Eq. (11). In contrast to retrieval precision, the decision vari-

able ignores the rank order of the retrieved cases but it takes into consideration the actual value of the similarity measure under consideration.

Table I shows the corresponding ROC areas achieved for each similarity measure based on the number k of the closest mass and normal templates retrieved from the knowledge database. Results are shown for several k values to highlight the general trends. For example, when $k=1$, the IT-CAD system is asked to make a decision using the one mass and one normal templates retrieved from the database as most similar to the query. In contrast, if $k=ALL$, the IT-CAD system is asked to make a decision using the whole knowledge database.

Using the mutual information, normalized mutual information, conditional entropy, and joint entropy as the similarity measure, the IT-CAD system achieved its highest ROC performance ($A_z=0.87 \pm 0.01$). Although not shown in Table I, the IT-CAD performed significantly better for the detection of malignant ($A_z=0.89 \pm 0.01$) than benign masses ($A_z=0.84 \pm 0.01$). The number of top mass and normal templates required for optimized performance depended on the similarity measure. Using mutual information, the system achieved its highest performance using as few as the top matched 50 mass and normal templates. Conditional entropy, normalized mutual information, and joint entropy required substantially more matched templates. The best ROC area index achieved by the IT-CAD scheme was significantly lower when using the Kullback-Leibler, Jensen, and arithmetic-geometric mean divergence measures ($A_z=0.77 \pm 0.01$). This performance was optimized with approximately 50 best matched mass and normal templates and deteriorated substantially as more inferior matches were included in the decision making process.

Since emphasis is typically place on operating at a high sensitivity level for breast cancer detection tasks, the impact of the eight similarity measures was also evaluated with respect to the partial ROC area index ${}_{0.90}A_z$. The overall trends remained the same. Specifically, MI, NMI, JOINT_H, and COND_H achieved significantly higher performance for malignant masses (${}_{0.90}A_z=0.57 \pm 0.03$) than the remaining measures (${}_{0.90}A_z=0.31 \pm 0.03$).

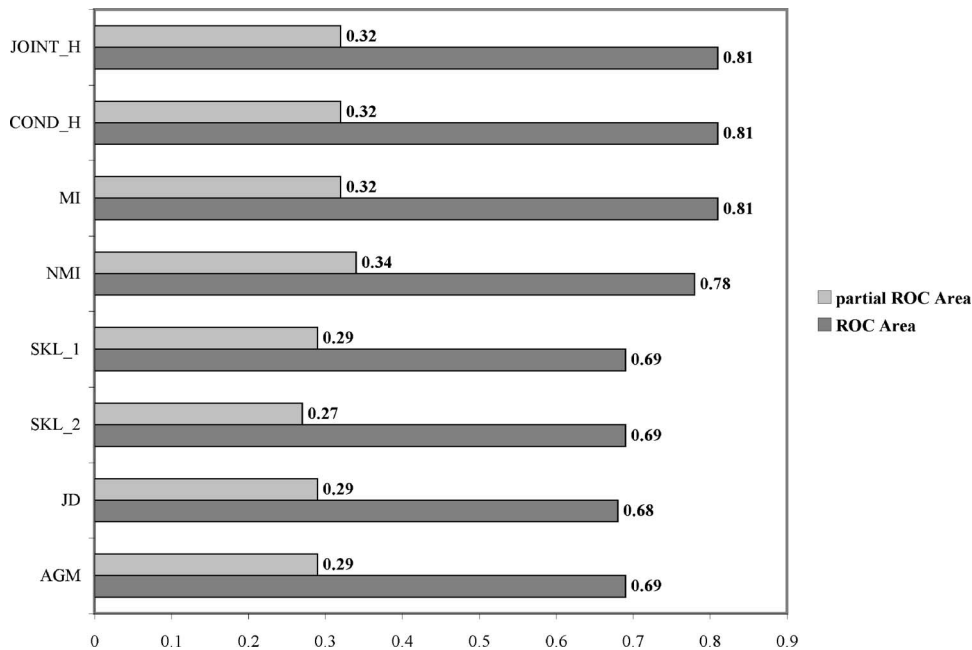


FIG. 3. Overall ROC area (± 0.02) and partial ROC area (± 0.04) indices achieved by the IT-CAD scheme for discrimination of true masses from suspicious yet normal ROIs depending on the image similarity measure.

C. Experiment 3: IT-CAD for false positive reduction

Finally, the IT-CAD scheme was validated for discriminating true masses from false positive mammographic regions. Figure 3 shows the overall ROC and partial ROC area indices achieved depending on the similarity measure. As the figure shows, the best performance of the IT-CAD scheme ($A_z = 0.81 \pm 0.02$) is significantly lower on Database 2 than what was previously observed on Database 1 ($A_z = 0.87 \pm 0.01$). The performance deterioration was expected because Database 2 represents a far more challenging detection task (mass vs. suspicious-looking normal ROIs) than Database 1 (mass vs. randomly chosen normal ROIs). However, the same overall trends prevailed. The same two groups of similarity measures emerged with distinctly different detection performance.

The IT-CAD detection performance on Database 2 was analyzed in more detail with NMI as the similarity measure. NMI is a more attractive choice than MI due to its bounded nature. It always ranges between 0 and 1 regardless of any possible preprocessing done on the ROIs. Three operating decision thresholds were selected using the partial ROC curves acquired in Database 1. The thresholds corresponded to three clinically relevant operating points: (a) 95% sensitivity, (b) 90% sensitivity, and (c) 85% sensitivity for malignant masses. Note that the decision thresholds for these three operating points were determined using Database 1 exclusively. Database 2 was used purely for testing.

Overall, the IT-CAD scheme had very robust detection performance in Database 2 when operating with the above decision thresholds. Operating at the desired 95% sensitivity decision threshold, the IT-CAD system detected 42/44 malignant masses present in Database 2 (95.7% sensitivity). At the 90% and 85% sensitivity operating thresholds, the scheme detected 40/44 malignant masses (90.9% sensitivity).

The contribution of the IT-CAD system for false positive reduction rate was also assessed at the same three operating points. Table II shows the false positive reduction rate for all false positives and for progressively more challenging ones; those remaining when the system operates at 3 FPs/image, 2 FPs/image, 1 FP/image, and 0.4 FP/image, respectively (at the expense of lower mass detection rate obviously). The IT-CAD scheme can effectively reduce about 50% of the false positive cues while detecting 90% of the malignant masses. Although the impact of the IT-CAD scheme deteriorates as the false positive cues become progressively more challenging, the scheme can still eliminate up to 17.5% of the false positive cues generated by the prescreening system that operates at a low 0.4 FP/image (while detecting 85% of malignant masses).

The results regarding retrieval precision in Database 2 were also consistent with what was observed in Database 1. The bin-by-bin similarity measures (SKL_1, SKL_2, JD, AGM) provided far more robust retrieval precision among all queries than the cross-bin similarity measures (JOINT_H, COND_H, MI, NMI). It is noted however, that the average retrieval precision for the false positive ROIs was consistently lower than that achieved for randomly chosen normal ROIs in Database 1 using the bin-by-bin similarity measures (0.60 vs 0.69).

IV. DISCUSSION

Assessment of image similarity is a critical step for the retrieval and diagnostic interpretation of medical images based on their content. The task involves two important decisions: (i) how to represent the image, and (ii) how to choose the most effective similarity measure for the specific image representation space and the particular medical task at hand. Typically, these decisions are made empirically using a

TABLE II. False positive reduction rate achieved by the IT-CAD scheme stratified according to the difficulty level of the cases. Results are shown for three malignant mass sensitivity [true positive fraction (TPF)] operating points. The IT-CAD scheme employed normalized mutual information as the similarity measure.

DIFFICULTY LEVEL OF FP CUES (No. of FPs/image)	% FP REDUCTION (Remaining average No. of FPs/image)		
	TPF=95%	TPF=90%	TPF=85%
ALL FPs (4 FPs/img)	29.8% (2.80)	45.9% (2.16)	52.0% (1.91)
75% most challenging FPs (3 FPs/img)	21.0% (2.37)	38.0% (1.86)	45.7% (1.63)
50% most challenging FPs (2 FPs/img)	14.0% (1.72)	29.5% (1.41)	36.0% (1.28)
25% most challenging FPs (1 FPs/img)	10% (0.90)	18.0% (0.82)	25.0% (0.75)
10% most challenging FPs (0.4 FPs/img)	5% (0.38)	7.5% (0.37)	17.5% (0.33)

labeled database. The size and comprehensiveness of the database usually determine how well the decisions generalize to new databases.

In the present study we investigated the retrieval performance and mass detection accuracy of eight information-theoretic (IT) image similarity measures for region-based analysis of mammograms. In contrast to feature-based similarity assessment techniques, the IT measures operate with image histograms without requiring image feature extraction. Thus, the image content is represented in terms of pixel intensity histograms. The IT similarity measures essentially compare the region-based histograms of two mammograms to determine how relevant they are. Specifically, this study focused on two groups of information theoretic measures: (i) bin-by-bin measures that compare only the contents of corresponding histogram bins (i.e., average KL divergence, maximum KL divergence, Jensen divergence, arithmetic-geometric mean divergence) and (ii) cross-bin measures (i.e., joint entropy, conditional entropy, mutual information, normalized mutual information) that incorporate the comparisons of the contents of noncorresponding bins.

The proposed image similarity measures were evaluated in the context of an interactive CAD system that is designed to provide evidence-based decisions regarding the presence of a malignant mass in mammographic locations that serve as queries for the system. These measures were evaluated in two different capacities: (i) for retrieval of diagnostically similar cases and for (ii) knowledge-based mass detection. Experiments were performed using two independent databases. The first database contained mammographic regions that depicted either a mass or normal breast parenchyma. This database was used for empirical comparison of the similarity measures based on a leave-one-case-out sampling scheme. The main conclusions drawn from using Database 1 were further validated on Database 2. The second database served as a clinically more challenging test bed because the nonmass mammographic regions it contained were already cued as highly suspicious for containing a mass by an in-house CAD system. Therefore, the additional validation experiment aimed to evaluate to what extent the information-

theoretic CAD analysis could improve upon the performance of existing CAD technology by providing evidence-based analysis of suspicious regions.

Our study clearly demonstrated two strong trends. First, bin-by-bin measures based only on the distance of the marginal histograms were more successful at achieving higher and more balanced average retrieval precision of cases with similar semantic content. High precision in the first few retrievals is critical for content-based image retrieval systems designed to display the top matches for visual evaluation by the CBIR user. On the other hand, cross-bin similarity measures that incorporate the joint histogram information were more successful for knowledge-based discrimination of masses from normal mammographic regions.

Based on the above observations, it seems reasonable to consider the ratio of retrieved masses over the total number of retrieved cases as a potential decision variable for the IT-CAD system. Basically, when a query case is presented for evaluation, the IT-CAD scheme retrieves the top K most similar cases. The prevalence of masses in the top retrievals is treated as a predictive variable for the presence of mass in the query image. This predictive variable is in essence similar to the odds ratio. If the query depicts a mass, then the above prevalence should be larger than if the query depicts normal breast parenchyma. Although not reported in this study, we explored this possibility with all similarity measures. As expected, the bin-by-bin similarity measures helped the IT-CAD scheme achieve a higher ROC area index than the cross-bin measures (0.74 ± 0.01 vs. 0.69 ± 0.01) for a low number of retrievals ($K < 30$). As more retrieved cases were considered, the ROC performance evened out between both groups of similarity measures. However, the ROC area index never exceeded the one achieved using the knowledge-based decision index [Eq. (11)] proposed in our study.

Finally, our study showed that the IT-CAD system can be effectively utilized as an add-on to existing detection schemes for false-positive reduction. Since the information-theoretic system follows a featureless-based image analysis, it appears to complement feature-based CAD schemes. Spe-

cifically, the IT-CAD system safely eliminated up to 17.5% of the most challenging false positive cues (those generated by prescreening the mammograms with a system that generates 0.4 FP/image) while still detecting 85% of the malignant masses. On a side note, the detection rate achieved for the benign masses was 90% (36/40).

To summarize, our study represents a critical step toward an interactive CAD system able to operate as an effective content-based image retrieval and knowledge-based mass detection system. The comparative analysis demonstrated that the choice of the similarity measure depends on the clinical task (retrieval vs. detection). No particular similarity measure emerges as the optimal choice for both tasks. While MI and NMI appear to be excellent choices for knowledge-based mass detection, they fail to provide robust retrieval precision across the two query classes for the top retrievals. Therefore, these measures are not suitable for CAD users who would like to view the top most relevant matches. In contrast, the bin-by-bin similarity measures such as Jensen divergence and Kullback-Leibler divergence achieved overall higher and more robust retrieval precision. However, these measures failed to reach the detection accuracy achieved by the cross-bin similarity measures. An interesting observation was that regarding retrieval precision, the cross-bin measures resulted in substantially lower pairwise correlation than the bin-by-bin measures. This finding suggests that MI, NMI, JOINT_H, and COND_H are good candidates for a possible fusion retrieval strategy. In fact, the newest trends in content-based image retrieval suggest that composite similarity measures may be more effective than single similarity measures. Our study certainly points toward that direction. For example, a composite strategy where a bin-by-bin similarity measure is used for initial retrieval of semantically similar cases while a cross-bin similarity measure is subsequently used for knowledge-based analysis of the retrieved cases appears to be a promising strategy to achieve simultaneously high retrieval precision and detection accuracy. We are currently investigating this idea.

One of the limitations of the present study design is that it assessed retrieval precision based on semantic, not visual content. This aspect is important for interactive CBIR-based CAD systems. It is possible that cross-bin measures may be more effective at capturing visual content than bin-by-bin measures. We plan to investigate this possibility in the future. In addition, we will investigate how beneficial the system is with mammographic regions that raise visual suspicion. The present study focused only on image locations marked as suspicious by another CAD algorithm. Analyzing image locations that are marked as suspicious by radiologists will determine the role of the IT-CAD system for reducing the interpretation, not perceptual, error associated with the diagnostic interpretation of mammograms.

From a theoretical point of view, the inherent limitation of the information-theoretic measures evaluated in this study is that they focus on the global histograms representation. Therefore, the localized spatial relationships among the image pixels are lost. This limitation has been addressed before

in the context of image registration. It has been proposed that taking into account the neighborhood of regions of corresponding image pixels may be a more effective strategy.^{44,45} The computational complexity and less dramatic than anticipated improvements of this approach have led researchers to seek simpler surrogate approaches. For example, Pluim, Maintz, and Ueirgever proposed multiplying the mutual information with an additional term that incorporates the local gradients of the two images in comparison.⁴⁶ Certainly advances made toward this direction in image registration may have significant implications for our CAD application as well.

In conclusion, this study represents a comprehensive step toward a framework of entropy-based, image similarity assessment for retrieval of diagnostically relevant images to support interactive, evidence-based diagnostic interpretation of mammograms.

ACKNOWLEDGMENTS

This work was supported by Grant No. R01 CA101911 from the National Cancer Institute and by Grant No. W81XWH-05-1-0293 from the Army Breast Cancer Research Program. We would like to thank Dr. Alan Baydush and Dr. David Catarious for providing guidance in the application of the prescreening CAD system to generate the suspicious mammographic regions.

^{a)}Electronic mail: georgia.tourassi@duke.edu

¹L. J. W. Burhenne *et al.*, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," *Radiology* **215**, 554–562 (2000).

²R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," *Radiology* **219**, 192–202 (2001).

³R. F. Brem *et al.*, "Improvement in sensitivity of screening mammography with computer-aided detection: A multiinstitutional trial," *AJR, Am. J. Roentgenol.* **181**(3), 687–693 (2003).

⁴R. F. Brem and J. M. Schoonjans, "Radiologist detection of microcalcifications with and without computer-aided detection: A comparative study," *Clin. Radiol.* **56**(2), 150–154 (2001).

⁵M. A. Helvie *et al.*, "Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection: Pilot clinical trial," *Radiology* **231**, 208–214 (2004).

⁶K. Moberg, N. Bjurstam, B. Wilczek, L. Rostgard, E. Egge, and C. Muren, "Computed assisted detection of interval breast cancers," *Eur. J. Radiol.* **39**, 104–110 (2001).

⁷P. M. Taylor, J. Champness, R. M. Given-Wilson, H. W. W. Potts, and K. Johnston, "An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms," *Br. J. Radiol.* **77**, 21–27 (2004).

⁸K. Hukkinen, T. Vehmas, M. Pamelu, and L. Kivisaari, "Effect of computer-aided detection on mammographic performance: Experimental study on readers with different levels of experience," *Acta Radiol.* **47**, 257–263 (2006).

⁹T. W. 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).

¹⁰D. Gur, H. Sumkin, H. E. Rockette, M. Ganott, C. Hakim, L. A. Hardesty, T. S. W. R. Poller, R. Shah, and L. Wallace, "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," *J. Natl. Cancer Inst.* **96**(3), 185–190 (2004).

¹¹P. Taylor and R. M. Given-Wilson, "Evaluation of computer-aided detection (CAD) devices," *Br. J. Radiol.* **78**, 26–30 (2005).

- ¹²R. L. Birdwell, P. Bandodkar, and D. M. Ikeda, "Computer-aided detection with screening mammography in a university hospital setting," *Radiology* **236**, 451–457 (2005).
- ¹³M. J. Morton, D. H. Whaley, K. R. Brandt, and K. K. Amrami, "Screening mammograms: Interpretation with computer-aided detection—Prospective evaluation," *Radiology* **239**(2), 375–383 (2006).
- ¹⁴R. M. Nishikawa and M. Kallergi, "Computer-aided detection, in its present form, is not an effective aid for screening mammography," *Med. Phys.* **33**(4), 811–814 (2006).
- ¹⁵E. A. Krupinski, "Computer-aided detection in clinical environment: Benefits and challenges for radiologists," *Radiology* **231**, 7–9 (2004).
- ¹⁶A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 1349–1380 (2000).
- ¹⁷H. Müller, A. Rosset, A. Garcia, J.-P. Vallée, and A. Geissbuhler, "Benefits of content-based visual data access in radiology," *Radiographics* **25**, 849–858 (2005).
- ¹⁸M. W. Vannier and R. M. Summers, "Sharing Images," *Radiology* **228**, 23–25 (2003).
- ¹⁹G. D. Tourassi, R. Vargas-Voracek, and C. E. Floyd, Jr., "Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information," *Med. Phys.* **30**, 2123–2139 (2003).
- ²⁰G. D. Tourassi and C. E. Floyd, Jr., "Computer-assisted diagnosis of mammographic masses using an information-theoretic image retrieval scheme with BIRADS-based relevance feedback," *Proc. SPIE* **5370**, 810–816 (2004).
- ²¹H. Alto, R. M. Rangayyan, and J. E. L. Desautels, "Content-based retrieval and analysis of mammographic masses," *J. Electron. Imaging* **14**(2), 023016 (2005).
- ²²I. El-Naqa, Y. Y. Yang, N. P. Galatsanos, R. M. Galatsanos, and M. N. Wernick, "A similarity learning approach to content-based image retrieval: Application to digital mammography," *IEEE Trans. Med. Imaging* **23**(10), 1233–1244 (2004).
- ²³M. O. Honda, P. M. A. Marques, and J. A. H. Rodrigues, "Content-based image retrieval in mammography: Using texture features for correlation with BI-RADS categories," *Proceedings of the 6th International Workshop on Digital Mammography*, Bremen, Germany, June 22–25, 401–403 (2002).
- ²⁴B. Zheng, A. Lu, L. A. Hardesty, J. H. Sumkin, C. M. Hakim, M. A. Ganott, and D. Gur, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," *Med. Phys.* **33**(1), 111–117 (2006).
- ²⁵C. Muramatsu, O. Li, K. Suzuki, R. A. Schmidt, J. Shiraishi, G. M. Newstead, and K. Doi, "Investigation of psychophysical measure for evaluation of similar images for mammographic masses: Preliminary results," *Med. Phys.* **32**(7), 2295–2304 (2005).
- ²⁶P. M. Azevedo-Marques and M. O. Honda, "Content-based image retrieval in mammography: Using texture features for correlation with BI-RADS categories," *Radiology* **221**(Suppl. S), 161–162 (2001).
- ²⁷C.-H. Wei, C.-T. Li, and R. Wilson, "A general framework for content-based medical image retrieval with its application to mammograms," *Proc. SPIE* **5748**, 134–143 (2005).
- ²⁸Y. Rubner, J. Puzicha, J. M. Bhumann, and C. Tomasi, "Empirical evaluation of dissimilarity measures for color and texture," *Comput. Vis. Image Underst.* **84**, 25–43 (2001).
- ²⁹S. Santani and R. Jain, "Similarity measures," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(9), 871–883 (1999).
- ³⁰T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
- ³¹J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes, *Medical Image Registration* (CRC, Boca Raton, FL, 2000).
- ³²W. M. Wells, P. V. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Med. Image Anal.* **1**, 35–51 (1996).
- ³³F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multi-modal image registration by maximization of mutual information," *IEEE Trans. Med. Imaging* **16**, 187–198 (1997).
- ³⁴W. Li, "Mutual information functions versus correlation functions," *J. Stat. Phys.* **60**, 823–837 (1990).
- ³⁵T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based feature distributions," *Pattern Recogn.* **29**(1), 51–59 (1996).
- ³⁶J. Puzicha, T. Hofmann, and J. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, June 17–19, 267–272 (1997).
- ³⁷M. Heath *et al.*, "Current status of the digital database for screening mammography," in *Digital Mammography* (Kluwer, Dordrecht, 1998). Available: <http://marathon.csee.usf.edu/Mammography/Database.html>.
- ³⁸D. M. Catarious, A. H. Baydush, and C. E. Floyd, Jr., "A mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: Preliminary results," *Proc. SPIE* **5032**, 111–119 (2003).
- ³⁹D. M. Catarious, A. H. Baydush, and C. E. Floyd, Jr., "Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system," *Med. Phys.* **31**(6), 1512–1520 (2004).
- ⁴⁰J. C. Russ, *The Image Processing Handbook* (CRC, Boca Raton, FL, 1995).
- ⁴¹D. Marr, *Vision* (Freeman, San Francisco, 1982).
- ⁴²D. Catarious, A. Baydush, and C. E. Floyd, Jr., "The influence of true positive detection definitions on the performance of a mammographic mass CAD system," *Med. Phys.* **30**(6), 1368–1368 (2003).
- ⁴³N. A. Obuchowski, "Receiver operating characteristic curves and their use in radiology," *Radiology* **229**, 3–8 (2003).
- ⁴⁴D. B. Russakof, C. Tomasi, T. Rohlfing, and C. R. Maurer, Jr., "Image similarity using mutual information of regions," *The Eighth European Conference on Computer Vision, ECCV*, Prague, Czech Republic, May 11–14, 596–607 (2004).
- ⁴⁵D. Rueckert, M. J. Clarkson, D. L. G. Hill, and D. J. Hawkes, "Non-rigid registration using higher-order mutual information," *Proc. SPIE* **3979**, 438–447 (2000).
- ⁴⁶J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Trans. Med. Imaging* **19**, 809–814 (2000).