A Similarity Study of Interactive Content-Based Image Retrieval Scheme for Classification of Breast Lesions

Hyun-chong CHO†, Member, Lubomir HADJISKI††, Berkman SAHINER†††, Heang-Ping CHAN†††, Chintana PARAMAGUL†††, Mark HELVIE†††, Alexis V. NEES†††, and Hyun Chin CHO††††, Nonmembers

SUMMARY To study the similarity between queries and retrieved masses, we design an interactive CBIR (Content-based Image Retrieval) CADx (Computer-aided Diagnosis) system using relevance feedback for the characterization of breast masses in ultrasound (US) images based on radiologists’ visual similarity assessment. The CADx system retrieves masses that are similar to query masses from a reference library based on six computer-extracted features that describe the texture, width-to-height, and posterior shadowing of the mass. The k-NN retrieval with Euclidean distance similarity measure and the Rocchio relevance feedback algorithm (RRF) are used. To train the RRF parameters, the similarities of 1891 image pairs from 62 (31 malignant and 31 benign) masses are rated by 3 MQSA (Mammography Quality Standards Act) radiologists using a 9-point scale (9=most similar). The best RRF parameters are chosen based on 3 observer experiments. For testing, 100 independent query masses (49 malignant and 51 benign) and 121 reference masses on 230 (79 malignant and 151 benign) images were collected. Three radiologists rated the similarity between the query masses and the computer-retrieved masses. Average similarity ratings without and with RRF were 5.39 and 5.64 for the training set and 5.78 and 6.02 for the test set, respectively. Average AUC values without and with RRF were, respectively, 0.86±0.03 and 0.87±0.03 for the training set and 0.91±0.03 and 0.90±0.03 for the test set. On average, masses retrieved using the CBIR system were moderately similar to the query masses based on radiologists’ similarity assessments. RRF improved the similarity of the retrieved masses.

key words: computer-aided diagnosis, relevance feedback, content-based image retrieval, breast lesion characterization

1. Introduction

One of the leading causes of death for women all over the world is breast cancer. In 2013, it is expected that approximately 234,580 new cases will be diagnosed and 40,030 deaths will occur in the United States [1]. Ultrasonography (US) has been shown to be a useful modality for characterizing breast masses as malignant or benign [2], [3]. In most breast imaging clinics, mammography is supplemented by sonography for the diagnosis of breast masses. However, sonography is most effective if it is used for real-time evaluation by an experienced interpreter [2]–[4], which in most clinical situations may not be practical. Besides this, between malignant and benign lesions there is considerable overlap in sonographic characteristics due to the heterogeneous appearance of breast cancer.

Computer-aided diagnosis (CADx) is a research area that has been explored to improve radiologists’ accuracy in characterizing malignant or benign lesions. Studies have shown that CADx, by providing a second opinion, can assist radiologists in the characterization of a given lesion or a region of interest based on radiology images [5], [6]. To characterize a lesion, radiologists not only interpret individual features of that lesion, but also use their recollection of clinically similar cases as references. Radiologists expand pattern recognition memory of specific appearances or patterns of lesions and their characterization. This process has been labeled “Aunt Minnie” [7], [8] in an analogy to human recognition of facial features. The current digital technologies for computing, networking, and database storage have empowered automated search for relevant and visually similar references in large image databases. The importance of image similarity for diagnostic decision making has activated wide research interest in the development of Content-Based Image Retrieval (CBIR) technology in medical imaging areas [9], [10]. Several groups are developing methods to incorporate CBIR approaches into image database systems [11]–[16].

We are developing a CBIR system for CADx of masses in US images. A CBIR system provides additional information to the radiologist by retrieving lesion images similar to the mass of interest (i.e., the query mass) from a reference library and presenting the known pathology of the retrieved masses as references to assist the radiologist in making diagnostic decisions about the query mass. If the reference library is statistically representative of the population and the prevalence is properly taken into account, the likelihood of malignancy of the query mass can be estimated by the CBIR system from the proportion of retrieved malignant and benign masses [17]. Increasing the performance of the CBIR system is an important research goal.

Developing the CBIR system is a complicated process for which many important issues have to be investigated, such as what the best similarity measure is for image retrieval, what the best method is for estimation of the likelihood of malignancy, and whether a given CBIR method can
retrieve lesions that are expected to be similar by radiologists. In this study, based on radiologists’ visual similarity assessment, we focus on the effect of relevance feedback on the similarity between a query and the retrieved masses in an interactive CBIR CADx system for characterization of breast masses in ultrasound (US) images.

2. Materials and Methods

2.1 Data Set

A data set was collected from the files of patients who had, with institutional review board (IRB) approval, undergone breast US imaging in the Department of Radiology at the University of Michigan. All US images from 250 patients with breast masses (96 malignant, 154 benign) were acquired by radiologists using a GE Logiq 700 scanner with an M12 linear array transducer. All masses had biopsy proven pathology. A total of 488 US images were selected for these masses. These images were randomly partitioned into two subsets: T1 and T2. The 258 images from 55 malignant and 74 benign masses were included in T1, and 230 images from 41 malignant and 80 benign masses were included in T2. The dataset selection procedure have been described in more detail before [16]. The average longest diameters of the malignant and benign masses were 12.5 and 7.2 mm, respectively (total range: 1.8-37.0 mm).

2.2 Feature Extraction and Selection

For the mass segmentation, an automated technique developed by Cui et al. [18] was applied. The method automatically calculated an initial contour based on a manually identified point approximately at the mass center. A two-stage active contour approach was used to iteratively refine the initial contour and to perform self-examination and correction of the segmentation result.

We extracted width-to-height, posterior shadowing, and texture features based on the computer segmentation. We defined the width-to-height ratio feature as a descriptor of the taller-than-wide shape of a sonographic mass, which is a good indication of malignancy [3]. The posterior shadowing feature also has been reported to be useful for differentiation of malignant and benign masses. We described the posterior shadowing as the normalized average gray-level difference between the interior of the segmented mass and the darkest posterior strip [19]. The texture features were extracted from the spatial gray-level dependence (SGLD) matrices. Six texture features were extracted: information measures of correlations 1 and 2, difference entropy, entropy, energy, and sum entropy. The mathematical definitions of these features can be found in [20]. The extraction and selection methods for these features have been described in greater detail previously [18]. A linear discriminant analysis (LDA) classifier [21] with stepwise feature selection was designed to classify the masses as malignant or benign using two-fold cross validation. Six features were selected by stepwise LDA, including the width-to-height feature, posterior shadowing feature, and 4 texture features.

2.3 Retrieval Method

Figure 1 provides a flow diagram of our CBIR scheme without relevance feedback. The masses on the US images from the reference database are segmented and the feature vectors that characterize the masses comprised a reference feature data set that is stored in the reference library. When a query mass is input to the CBIR system to search for similar masses, the system first extracts the same feature vector as that of the reference library from this query mass. Using the similarity measure, the similarity scores between the feature vector of the query sample and those in the reference library are calculated. The system ranks the similarity scores and retrieves the reference masses that are most similar to the query sample.

Various similarity measures and retrieval methods can be used in a CBIR system [16]; the feature vectors are designed based on the characteristics of the samples in a given application or situation. In this study, our CBIR system automatically extracts the width-to-height ratio, the posterior shadowing feature, and texture features from the spatial gray level dependence matrices; it then retrieves masses that are similar to a query mass from a reference library based on the Euclidean distance similarity measure (ED). The k-Nearest Neighbor (k-NN) algorithm is used to retrieve k reference masses that have the highest similarity scores with the query mass [16].

2.4 Relevance Feedback CBIR

A relevance feedback CBIR CADx system is shown in Fig. 2. The system constitutes the CBIR system shown Fig. 1, with an adaptation module that performs relevance feedback.

The adaptation module integrates the Rocchio relevance feedback algorithm (RRF) [22], [23].

\[
\bar{Q}_m = \alpha \bar{Q}_0 + \beta \sum_{D_j \in D_p} \bar{D}_j - \gamma \sum_{D_k \in D_n} \bar{D}_k
\]  

(1)

where: \( \bar{Q}_m \) is the modified query vector, \( \bar{Q}_0 \) is the original
query vector, \( \vec{D}_j \) is relevant image vector, \( \vec{D}_k \) is non-relevant image vector, \( D_r \) is the set of relevant image vector, \( D_{nr} \) is the set of non-relevant image vector, \( \alpha \) is original query weight, \( \beta \) is relevant image weight, \( \gamma \) is non-relevant images weight, and \( \alpha, \beta, \text{and } \gamma \) are balancing weights.

First, as in a regular CBIR system, for the input query mass, the CBIR system retrieved the top \( k \) most similar masses from the reference library using the ED similarity measure and the computer-extracted features that describe the properties of a mass. Then, to obtain the relevance feedback, the user (i.e., a radiologist, in our study) rated the \( k \) image pairs (one query mass, \( k \) retrieved masses) using a nine-point similarity scale (1 = very dissimilar, 9 = very similar). From the radiologist’s ratings and a threshold, the pairs were separated into categories of relevant or irrelevant. The RRF algorithm generates a virtual query feature vector by modifying the vector of the original query (Fig. 2). The virtual query feature vector is calculated as a combination of the original query vector, the vectors of the retrieved relevant masses, and the vectors of the retrieved irrelevant masses. Scaling coefficients are applied for each of the vector groups in order to move the virtual query vector in the direction of the relevant samples and away from the irrelevant samples. The scaling coefficients (i.e. \( \alpha, \beta, \text{and } \gamma \)) are the parameters of the RRF algorithm that need to be trained. After the virtual query vector is calculated, the CBIR system retrieves from the reference library the top \( k \) masses most similar to the virtual query mass. The radiologist again rates the latterly retrieved \( k \) image pairs. This process can continue for multiple iterations.

2.5 Similarity Evaluation by Radiologists

Based on radiologists’ visual similarity assessments, we evaluated the similarity between queries and retrieved masses using the CBIR CADx system. Three MQSA radiologists independently rated the similarity between the query masses and the CBIR-retrieved masses using a 9-point similarity scale (1 = very dissimilar, 3 = quite dissimilar, 5 = some degree of resemblance, 7 = quite similar, and 9 = very similar). The similarity ratings 2, 4, 6, and 8 are intermediate ratings.

2.6 Training RRF CBIR System

For the training set, from the T2 set, we randomly selected 31 images from 41 malignant masses, and 31 from 80 benign masses. To train the RRF parameters, the similarities of 1891 image pairs from the 62 (31 malignant and 31 benign) masses were rated by 3 MQSA radiologists using the 9-point scale, as described before. During training, with a leave-one-out method, the 5 most similar masses to each query mass were retrieved from the reference library using Euclidean distance as similarity measure. Then, the radiologists’ similarity ratings were used by the RRF to retrieve another 5 masses for the same query. The best RRF parameters were chosen based on the simulated observer experiments, which used the average values of the three radiologists’ ratings of the 1891 pairs for retrieval and relevance feedback.

2.7 Testing RRF CBIR System Performance

To evaluate the performance of the RRF CBIR system, we used 100 (49 malignant and 51 benign) images of 100 query masses randomly selected from the T1 set, which, as the test set, is independent of the RRF training set and the reference library. The reference library (T2 set) contains 121 masses on 230 (79 malignant and 151 benign) images.

For each query mass, the 5 most similar masses (\( k=5 \)) were retrieved from the reference library first without RRF and then with RRF. One RRF iteration was applied. Therefore, a total of 1000 \( (100 \times 5 \times 2) \) pairs of queries and retrieved masses were formed and evaluated without and with RRF by each radiologist in the observer experiment. The mass pairs were shown to the radiologists one pair at a time. Three MQSA radiologists were participants in the observer study. The radiologists rated the similarity between the query masses and the RRF CBIR-retrieved masses using the 9-point similarity scale. Due to the constraint of available time for the radiologists, the determination of the number of observers, the number of RRF iterations, the number of query masses, and the number of retrieved masses (\( k \)) was a compromise among these factors in order to complete the similarity study within a reasonable amount of time.

2.8 Evaluation Methods

2.8.1 Evaluation of Classification Performance of CBIR

Based on the known pathology of the retrieved masses, the proportion of the retrieved malignant masses can be used as an estimation of the malignancy score of the query masses as follows:

\[
p = \frac{1}{k} \sum_{i=1}^{k} b_i
\]

where \( k \) is the number of retrieved masses and \( b_i \) is a binary index from the known pathology database in the reference library indicating whether a retrieved mass is malignant (1) or benign (0). The malignancy scores were then analyzed using the receiver operating characteristics (ROC) methodology and the area under the ROC curve (AUC) was calculated. This analysis was performed for the top 5 most similar
retrieved masses ($k = 5$).

2.8.2 Similarity Evaluation by Radiologists

We evaluated the average similarity ratings between the query and the retrieved masses using the RRF CBIR CADx system without and with RRF based on radiologists’ visual similarity assessments for all retrieved masses under each condition and also over all three radiologists.

3. Results

3.1 Similarity on RRF Training Set

Three radiologists rated the similarity of all possible pairs (1891 image pairs) from the 62 masses. Table 1 shows the results of the similarity ratings. One of radiologists (R1) set 8 as the highest similarity score in the training set and the other two (R2 and R3) set 9 as the highest similarity score. The average of the three radiologists’ similarity ratings and the standard deviation are 4.29 and 1.66, respectively.

Figure 3 shows the histograms of the similarity ratings from the 1891 pairs. Figure 3 (a), (b), and (c) are histogram of all pairs rated by R1, R2 and R3, respectively. Figure 3 (d) shows the average histogram for all pairs by R1, R2, and R3; many pairs are at similarity scale 3 (i.e. quite dissimilar), 4, and 5 (i.e. some degree of resemblance). From Fig. 3 (e) and 3 (f), malignant pairs can be seen to have a higher number of pairs of high similarity than is the case for benign pairs.

3.2 Training the RRF Parameters

Cho et al. [16] applied 4 similarity measures (Euclidean Distance, Cosine Measure, Linear discriminant Analysis, and Bayesian Neural Net) for the similarity evaluation study. Since Euclidean Distance produced good results and is the standard method, this measure is used in this study.

The RRF parameters ($\alpha$, $\beta$, and $\gamma$) are the constants that regulate the weight each component has in the formation of a new query. Typically, $\alpha = 1$ and $\beta + \gamma = 1$ [24]. We set these values as $\alpha = 1$, $\beta = 0.65$, and $\gamma = 0.35$. These parameters are set to retrieve most visually similar images in the training set. Tables 2 and 3 show the average similarity and the average $AUC$ values assigned by the three radiologists from the training set while changing the similarity scale as

<table>
<thead>
<tr>
<th>Radiologist</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>4.48</td>
<td>1.70</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>R2</td>
<td>3.89</td>
<td>2.02</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>R3</td>
<td>4.49</td>
<td>2.16</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Average</td>
<td>4.29</td>
<td>1.66</td>
<td>1</td>
<td>8.67</td>
</tr>
</tbody>
</table>

Table 2 Average similarity of three radiologists without and with relevance feedback changing threshold values in the training set. Threshold value which decides relevant or irrelevant is one of similarity scale (1-9).

<table>
<thead>
<tr>
<th>Average Similarity</th>
<th>Threshold = 4</th>
<th>Threshold = 5</th>
<th>Threshold = 6</th>
<th>Threshold = 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Feedback</td>
<td>5.39</td>
<td>5.39</td>
<td>5.39</td>
<td>5.39</td>
</tr>
<tr>
<td>With Feedback</td>
<td>5.46</td>
<td>5.64</td>
<td>5.64</td>
<td>5.44</td>
</tr>
</tbody>
</table>

Table 3 Average $AUC$ of three radiologists without and with the relevance feedback changing threshold values in the training set. Threshold value which decides relevant or irrelevant is one of similarity scale (1-9).

<table>
<thead>
<tr>
<th>Average $AUC$ Value</th>
<th>Threshold = 4</th>
<th>Threshold = 5</th>
<th>Threshold = 6</th>
<th>Threshold = 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Feedback</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>With Feedback</td>
<td>0.87</td>
<td>0.86</td>
<td>0.87</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Fig. 4  Malignant query mass and five masses \((k = 5)\) retrieved by our CBIR scheme before feedback: (a) malignant query mass, (b) first retrieved mass, (c) second retrieved mass, (d) third retrieved mass, (e) fourth retrieved mass, (f) fifth retrieved mass. The biopsy results of (a), (c), (d), (e), and (f) are malignant and result of (b) is benign. The similarity ratings by one of three radiologists between the query mass and the retrieved masses are: (b) 2; (c) 3; (d) 8; (e) 9; (f) 9.

Fig. 5  Identical malignant query mass to that in Fig. 4 and five masses \((k = 5)\) retrieved by our CBIR scheme after feedback: (a) malignant query mass, (b) first retrieved mass, (c) second retrieved mass, (d) third retrieved mass, (e) fourth retrieved mass, (f) fifth retrieved mass. The biopsy results of (a), (b), (c), (d), (e) and (f) are malignant. The similarity ratings by one of three radiologists between the query mass and the retrieved masses are: (b) 8; (c) 9; (d) 8; (e) 9; (f) 6.

3.3 Testing RRF CBIR System Performance

Examples of similarity evaluations by one of the radiologists are shown on Figs. 4 and 5, with and without the feedback for the malignant query mass.

Figure 6 shows the similarity ratings and \(AUC\) values of the area under receiver operating characteristic (ROC) curves for top 5 retrieved masses and application of Rocchio Relevance Feedback (RRF) algorithm: (a) similarity ratings from training set, (b) similarity ratings from test set, (c) \(AUC\) values from training set, and (d) \(AUC\) values from test set.

In this study, our focus is to design an interactive CBIR CADx system to retrieve more similar masses. Although radiologists have their own similarity rating scales, our system retrieves more similar masses after the feedback. Table 1 and Fig. 3 shows similarity scores from three radiologists. The average number of changed masses by two methods (without and with feedback) is 2.57.

In Fig. 6(d), the \(AUC\) values with RRF for the test set are decreased from those without RRF for two radiologists. The difference of R2 is relatively small comparing those of R1. Since the main goal of this research is to retrieve most visually similar masses, the RRF parameters are chosen based on the highest similarity results in the training set. This may cause the decreased \(AUC\) values in the testing set. The average decreased \(AUC\) value is 0.006. The ultimate criterion for a CADx system is the improvement in the performance of the radiologists when they are aided by the CADx system. The evaluation of interactive CBIR CADx system performance is a nearly new area, and the trade-offs between the performances of the stand-alone system for retrieval and classification as they are related to this ultimate criterion are not yet known. Future observer studies may tell us about the relative importance of these two performance criteria.
Table 4  Average similarity of three radiologist without and with relevance feedback change k values in the training set. Threshold value for the feedback is the scale 6 and RRF parameters are $\alpha = 1$, $\beta = 0.65$, and $\gamma = 0.35$.

<table>
<thead>
<tr>
<th>Average Similarity</th>
<th>$k = 3$</th>
<th>$k = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Feedback</td>
<td>5.40</td>
<td>5.39</td>
</tr>
<tr>
<td>With Feedback</td>
<td>5.60</td>
<td>5.64</td>
</tr>
</tbody>
</table>

Table 5  Average AUC of three radiologist without and with relevance feedback change k values in the training set. Threshold value for the feedback is the scale 6 and RRF parameters are $\alpha = 1$, $\beta = 0.65$, and $\gamma = 0.35$.

<table>
<thead>
<tr>
<th>Average AUC Value</th>
<th>$k = 3$</th>
<th>$k = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Feedback</td>
<td>0.83</td>
<td>0.86</td>
</tr>
<tr>
<td>With Feedback</td>
<td>0.86</td>
<td>0.87</td>
</tr>
</tbody>
</table>

The number of retrieved masses, $k$, can be an important parameter for the performance. We set $k = 5$ because 5 is the odd number for the majority vote and we need to consider the constraint of available time for the radiologists. Tables 4 and 5 show average similarity and AUC of three radiologist without and with relevance feedback change $k$ values in the training set. The $k = 5$ shows the better results of average similarity and AUC values in the training set with feedback. Our similarity study has limitations. In the cases of without feedback, the differences between $k = 3$ and $k = 5$ are caused by the differences between the number of retrieved reference masses in the evaluation. Three radiologists rate the similarity of a query mass to the top five ($k = 5$) retrieved masses. The total number of query masses are 100. Therefore, each radiologist performed up to 1000 (5×100×2) readings. Although the total number of readings was fairly large, the number of query masses and the number of readings for each mass were still small. In a future study, we will increase the number of observers in order to achieve more robust results. Increasing $k$ may also produce more reliable results; however, we have to carefully choose the value of $k$ in order to avoid excessive reading times for the radiologists. Finding a good balance among the number of observers, the number of query masses, and the $k$ value will be studied in the future.

In this study, we include orthogonal views of the same mass in the image library due to the limited data set available. It is possible that the retrieved images are orthogonal views of the same mass. Since our retrieval system is the lesion-based approach, not the patient-based, this is not big issue. However, future studies will be needed to further investigate this issue.

5. Conclusions

A well-developed CBIR CADx system can make CADx output more interpretable to radiologists, which potentially may improve radiologists’ characterization of breast masses in ultrasound images. We are developing a relevance feedback CBIR CADx system to assist radiologists in differentiating between benign and malignant masses in ultrasound breast images.

This preliminary study showed that, based on radiologists’ similarity assessments, masses retrieved by the CBIR system were, on average, moderately similar to the query masses. RRF improved the similarity of the retrieved masses but did not have a significant effect on the estimation of the malignant or benign ratings of the masses.

Future work should include expanding the reference library, verifying the results of this study by applying the RRF CBIR CADx system to a large test data set, exploring different adaptive algorithms, increasing the number of radiologists, and studying the effect of radiologists’ inter-observer and intra-observer variability on their performance with the relevance feedback CBIR CADx system.

Acknowledgments

This work is supported by USPHS Grant Nos. CA 118305 and 2014 Research Grant from Kangwon National University.

References


Mark Helvie is a professor of radiology, division director of breast imaging in University of Michigan, USA and a fellow of the Society of Breast Imaging. He is fellowship trained in breast imaging, CT, and ultrasound. A graduate of Duke University and the University of North Carolina, he took his residency training at the University of Michigan, becoming board certified in radiology and internal medicine. His primary research interest is breast cancer. He has authored and co-authored more than 100 publications, numerous grants and is an examiner for the American Board of Radiology.

Alexis V. Nees is an associate professor of radiology in University of Michigan, USA. Her specialties are Breast Imaging, Abdominal Radiology, and Radiology. Her clinical interests are Breast Imaging, ultrasound and computed tomography.

Hyun Chin Cho received the M.S. and Ph.D. degrees in Internal Medicine from Gyeongsang National University School of Medicine of Jinju, South Korea in 2008 and 2014, she was a Fellow at Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea from 2009 to 2010. She was an Assistant Professor at Department of Internal Medicine, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine, Changwon, South Korea from 2011 to 2015. She is currently an Clinical Assistant Professor at Department of Internal Medicine & Institute of Health Sciences, Gyeongsang National University School of Medicine, Jinju, South Korea.