Learning of Perceptual Similarity From Expert Readers for Mammogram Retrieval

Liyang Wei, *Member, IEEE*, Yongyi Yang, *Senior Member, IEEE*, Miles N. Wernick, *Senior Member, IEEE*, and Robert M. Nishikawa

***Abstract*—Content-based image retrieval relies critically on the use of a computerized measure of the similarity (i.e., relevance) of a query image to other images in a database. In this work, we explore a superivised learning approach for retrieval of mammogram images, of which the goal is to serve as a diagnostic aid for breast cancer. We propose that the most meaningful measure is one that is designed specifically to match that perceived by the radiologists in their interpretation of mammogram lesions. In our approach, we model the notion of similarity as an unknown function of the image features characterizing the lesions, and use modern machine-learning algorithms to learn this function from similarity scores collected from radiologists in reader studies. This approach is evaluated using data collected from an observer study with a set of clinical mammograms. Our results demonstrate that the proposed machine learning approach can be used to model the notion of similarity as judged by expert readers in their interpretation of mammogram images and that it can outperform alternative similarity measures derived from unsupervised learning.**

***Index Terms*—Content-based image retrieval, mammogram, multidimensional scaling, perceptual similarity, similarity measure, supervised learning.**

I. Introduction

The PURPOSE of content-based image retrieval (CBIR) is to choose images from a database on the basis of image content, such as color, texture, object shape, etc. [1], such that the retrieved information is most relevant or similar to a user's query. In recent years, there have been growing interests in development of CBIR for medical images [2], [3]. In a clinical decision-making process, CBIR can be used as a diagnostic aid by displaying cases of similar visual appearance with known pathologies. It can also be useful as a training tool for medical students in education, follow-up studies and for research purposes.

CBIR for medical images is challenging due to the complexity in their content in relation to the disease conditions. As a consequence, many of the useful image features in traditional CBIR are no longer adequate. For example, global image features (such as gray-scale histogram) would not be salient for describing the characteristics of pathological regions or lesions which are typically localized in the images [2]. In such a case, it is important to derive quantitative features that correlate well with the anatomical or functional information perceived as important for diagnostic purposes by the physicians.

Breast cancer remains to be a leading cause of death among women in developed countries. Currently mammography is the dominant method for detection of breast cancer. But it is still far from being perfect. The high sensitivity of screening mammography is compromised by its low specificity to benign lesions, which often appear mammographically similar to malignant lesions [4], [5]. Application of CBIR to mammography has been pioneered by Swett [6], where a rule-based expert system was developed to display chest radiographs from a library of images. In [7], a retrieval system for mammograms was developed based on tumor shapes. Kuo [8] studied CBIR for breast tumor based on sonogram. In our previous work [9], we developed a CBIR system for retrieving similar mammogram images from a database, of which the goal is to serve as a diagnostic aid to radiologists in their interpretation of mammograms. We conjecture that by presenting perceptually similar mammograms with known pathology to the one being evaluated, the radiologists could reach better informed decision in their diagnosis.

A key in the development of CBIR is the definition of the measure used for describing the similarity between a query and the images in a database. In [9], we proposed a supervised learning approach in which the similarity measure between two lesion images was modeled by machine learning from examples collected in human-observer studies. The rationale behind this approach is that the similarity metric must conform closely to the notion of similarity used by radiologists when they interpret the mammograms and that simple, mathematical distance metrics developed in the context of general-purpose image retrieval, e.g., Euclidean distance [10], Mahalanobis distance [11], and the earth mover's distance [12], may not be adequate. In [9], the similarity between a pair of mammograms was based on the perceptual similarity of their clustered microcalcifications (MCs).

MCs are calcium deposits of very small dimension and appear as granular bright spots in mammograms. MCs can be an important early indicator of breast cancer in women. Though commonly seen on mammograms, MCs are often difficult to di-

Manuscript received April 15, 2008; revised October 15, 2008. Current version published February 19, 2009. This work was supported in part by NIH/NCI Grant CA89668. R. M. Nishikawa is a shareholder in Hologic, Inc, Bedford, MA, and also receives research support and royalties. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jinshan Tang.

L. Wei is with the Department of Biomedical Engineering, Illinois Institute of Technology, Chicago, IL 0616 USA.

Y. Yang is with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616 USA (e-mail: yy@ece.iit.edu).

M. N. Wernick is with the is with the Department of Electrical and Computer Engineering and the Department of Biomedical Engineering, Illinois Institute of Technology, Chicago, IL 60616 USA.

R. M. Nishikawa is with the Department of Radiology, University of Chicago, Chicago, IL 60637 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2008.2011159

agnose. This greatly compromises the quality of radiologists' biopsy recommendations, which is an important issue in breast cancer diagnosis [13]. In [9], for the purpose of demonstrating feasibility, the notion of similarity between two lesion images was based on only the geometric distribution of the MCs, and the image features of individual MCs were ignored. Our goal at the time was first to demonstrate that the notion of similarity could be modeled by a machine learning approach. Encouraged by the success in [9], we now extend this approach by using similarity data collected from clinical expert readers in their interpretation of MCs, where the image features of individual MCs are considered. This will bring us one step closer toward the eventual development of a clinical diagnostic aid. The added difficulty with individual MCs is that their visual appearance can be quite subtle and may vary a great deal in mammograms. This to a large extent contributes to significant inter-observer variations in interpretation of clustered MCs [14]. Therefore, it is important to investigate to what degree the judgment on perceptual similarity by experts can be modeled by the image features of the individual MCs.

In this study, we use a set of data collected from a group of expert readers to investigate the feasibility of a machine learning approach for modeling perceptual similarity. We also compare the supervised learning approach with alternative similarity measures based on unsupervised learning. Furthermore, the multidimensional scaling (MDS) technique is used as a perceptual evaluation tool for displaying the retrieved results.

II. Methodology

A. Supervised Learning-Based Similarity Measure

As described in [9], the notion of similarity is modeled as a nonlinear function of the image features in a pair of mammogram images containing microcalcification clusters (MCCs). Specifically, let vectors $\mathbf{u}$ and $\mathbf{v}$ denote the features of two MCCs at issue. We use the following regression model for their similarity coefficient (SC):

$$\mathrm{SC}(\mathbf{u},\mathbf{v}) = f(\mathbf{u},\mathbf{v}) + \zeta \tag{1}$$

where $f(\mathbf{u},\mathbf{v})$ is a function determined using support vector machine (SVM) learning [15], and ζ is the modelling error. The similarity function $f(\mathbf{u},\mathbf{v})$ in (1) is trained using data samples collected in an observer study. For convenience, we denote $f(\mathbf{u},\mathbf{v})$ by $f(\mathbf{x})$ with $\mathbf{x} = [\mathbf{u}^T, \mathbf{v}^T]^T$.

Assume that we have a set of N training samples, denoted by $Z = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where y_i denotes the user similarity score for the MCC pair denoted by $\mathbf{x}_i, i = 1, 2, \ldots, N$. The regression function $f(\mathbf{x})$ is written in the following form:

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + b \tag{2}$$

where $\Phi(\mathbf{x})$ is a mapping implicitly defined by a so-called kernel funciton which we introduce below. The parameters $\mathbf{w}$ and b in (2) are determined through minimization of the following structured risk:

$$R(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} L_\varepsilon(\mathbf{x}_i) \tag{3}$$

where $L_\varepsilon(\bullet)$ is the so-called ε-insensitive loss function, which has the property that it does not penalize errors below the parameter ε, defined as

$$L_\varepsilon(\mathbf{x}) = \begin{cases} |y - f(\mathbf{x})| - \varepsilon, & \text{if } |y - f(\mathbf{x})| \geq \varepsilon \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

As with the case of classification, the constant C in (3) determines the trade-off between the model complexity and the training error. In this study the Gaussian radial basis function is used for the SVM kernel function $K(\cdot,\cdot)$, where $K(\cdot,\cdot) = \Phi^T(\mathbf{x})\Phi(\mathbf{x})$. The regression function $f(\mathbf{x})$ in (2) is characterized by a set of so-called support vectors [15]

$$f(\mathbf{x}) = \sum_{j=1}^{Ns} \gamma_j K(\mathbf{x}_j, \mathbf{x}) + b \tag{5}$$

where $\mathbf{x}_j$ are the support vectors, and N_s is the number of the support vectors.

To model the similarity between two feature vectors, we want to learn a symmetric function satisfying $f(\mathbf{u},\mathbf{v}) = f(\mathbf{v},\mathbf{u})$, i.e., the notion of similarity is commutative. This can be achieved by duplicating the training image pairs, i.e., first with $(\mathbf{u},\mathbf{v})$ and then with $(\mathbf{v},\mathbf{u})$. We can explicitly enforce this property in the SVM cost function as

$$R(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} L_\varepsilon(\mathbf{x}_i) + C\sum_{i=1}^{N} L_\varepsilon\left(\mathbf{x}_i^s\right). \tag{6}$$

Here, $y(\mathbf{x}_i) = y(\mathbf{x}_i^s), \mathbf{x}_i = (\mathbf{u}_i^T, \mathbf{v}_i^T)^T . \mathbf{x}_i^s = (\mathbf{u}_i^T, \mathbf{v}_i^T)^T$.

With this formulation, the SVM training algorithm yields the global optimum of a symmetric Lagrangian. The resulting regression function can be written as

$$f(\mathbf{x}) = \sum_{j=1}^{Ns} \gamma_j \left[K(\mathbf{x}_j, \mathbf{x}) + K\left(\mathbf{x}_j^s, \mathbf{x}\right)\right] + b. \tag{7}$$

That is, if a training sample $\mathbf{x}_j$ is a support vector, i.e., $|y_j - f(x_j)| \geq \varepsilon$, then its symmetric sample $\mathbf{x}_j^s$ is also a support vector and $\gamma_j = \gamma_j^s$. This will ensure that the solution is symmetric: $f(\mathbf{x}) = f(\mathbf{x}^s)$. A detailed proof of this is given in Appendix A.

B. Similarity Measure From Expert Observers

1) Data Set: The proposed retrieval framework was developed and tested using a database of mammogram images provided by the Department of Radiology at the University of Chicago. The database consists of a total of 200 different mammogram images of dimension 1024 × 1024 (some are 512 × 512) from 104 patients with known pathology (46 malignant, 58 benign), digitized with a spatial resolution of 0.1 mm/pixel and 10-bit grayscale. All these images contain microcalcification clusters (MCCs). The MCCs in each image have been identified by expert radiologists.

2) Human Observer Study: With 200 images we can form as many as 19 900 different image pairs. However, it would be too time-consuming (and also unnecessary) to score all of them in an observer study. Instead, we selected a subset consisting of a total of 600 representative image pairs for the observer study. These image pairs were selected using the following procedure

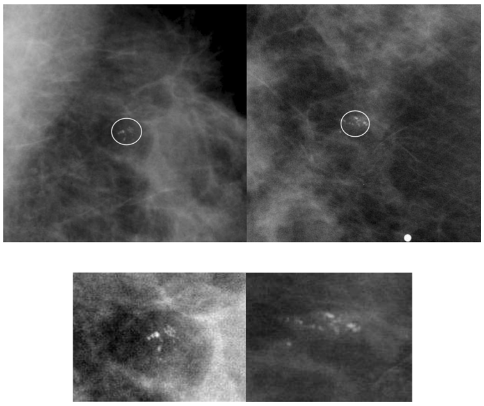

Fig. 1. Anchor image pair $(\mathrm{SC} = 9)$: original view (top) and magnified view (bottom).

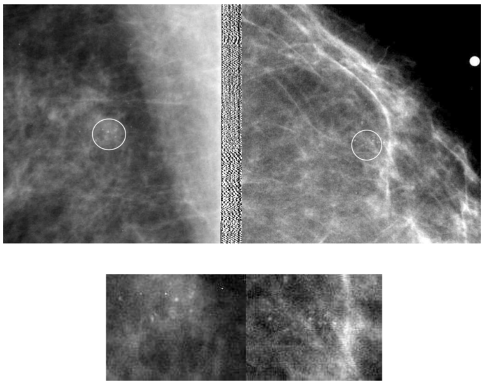

Fig. 2. Anchor image pair $(\mathrm{SC} = 7)$: original view (top) and magnified view (bottom).

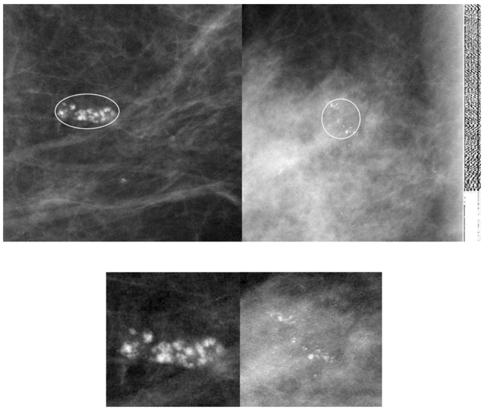

Fig. 3. Anchor image pair $(\mathrm{SC} = 5)$: original view (top) and magnified view (bottom).

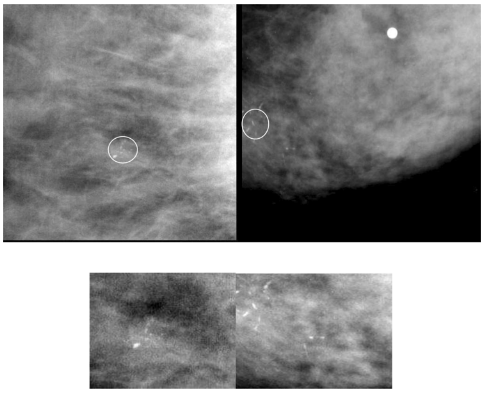

Fig. 4. Anchor image pair $(\mathrm{SC} = 3)$: original view (top) and magnified view (bottom).

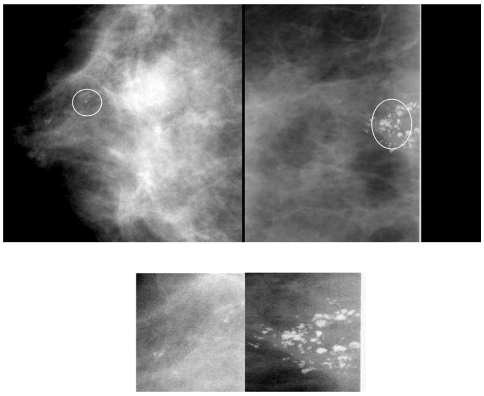

Fig. 5. Anchor image pair $(\mathrm{SC} = 1)$: original view (top) and magnified view (bottom).

so that they represent the spectrum of the image pairs in terms of similarity. First, the 200 images in the database were partitioned into ten different groups by the k-means method based on the features of their MCCs. The features used were the eight image features in [13] which were demonstrated to have high discriminating power for cancer diagnosis. Next, a total of 300 intra-group pairs were randomly selected, of which each pair was formed by images from a common group. Finally, a total of 300 inter-group pairs were randomly selected, of which each pair was formed by images from two different groups. In both cases, the probability that an image was selected was proportional to the size of the group that it was in. Conceivably, an intra-group pair is more likely (though not definitive) to be similar as their image features are closer in distance; in contrast, an inter-group pair is less likely to be similar.

The observer study was carried out by a panel of six expert observers, who scored the 600 pairs based on their perceptual similarity using a scale from 0 (most dissimilar) to 10 (most similar). It consisted of the following different sessions: 1) a "pre-calibration" session and 2) individual scoring sessions. The goal of the pre-calibration was to establish a consensus among the observers on a uniform measure of the perceptual similarity. In the pre-calibration session, five "anchor" image pairs were used to define the rating scale (their scores were 1, 3, 5, 7, 9, respectively). We show in Fig. 1 through 5 five anchor pairs with scores 9, 7, 5, 3, and 1, respectively. As can be seen in Fig. 1, while the images may look very different, their MC features can be very similar to an expert reader. In the individual scoring sessions, each observer scored the 600 image pairs separately. In addition, for the purpose of evaluating intra-observer consistency, each observer also scored 30 additional image pairs; these image pairs were presented in a random order, and each pair was scored twice by the same observer.

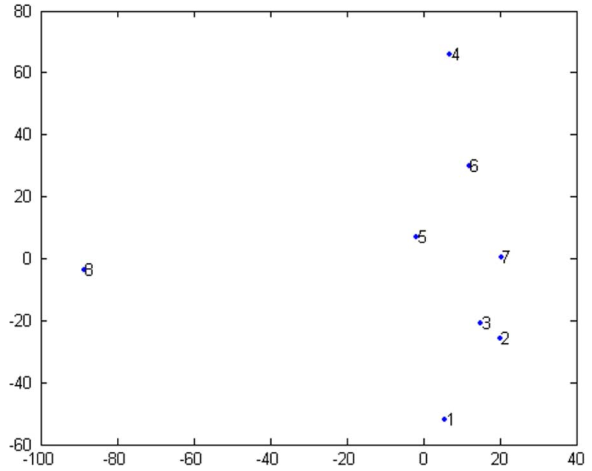

Fig. 6. MDS plot of six expert observers. No. 7 is the average of the six observers; No. 8 is a random observer.

Statistical analyses were conducted to analyze both intra- and inter-observer consistencies to insure the integrity of the observer data. For intra-observer consistency, Spearman's rank correlation method [16] was used to analyze the two sets of similarity scores from each observer on the 30 image pairs. It was found that there was statistically significant consistency within each observer with their p-values all below 0.05. For inter-observer consistency, Kendall's coefficient of concordance W [16] was computed, and there was statistically significant agreement among the observers with p-value less than 0.0001. Based on these analyses, we selected the four observers with the highest intra-observer consistency (2, 3, 5, 6), and averaged their similarity scores for each of the 600 image pairs. The resulting scores were used to form training samples for the SVM. We also show in Fig. 6 a multidimensional scaling (MDS) [17] plot of the six expert observers based on their scores. No. 7 is the average of the six observers and No. 8 is based on a random observer for which random scores were assigned. We could clearly see that the scores from the six observers are close to each other and far away from random scores.

C. Similarity Training and Feature Selection

In our previous work [9], a set of ten features was used based on the geometric distribution of the MCs in a cluster. However, as we mentioned earlier, image features of individual MCs are very important for diagnosis of clustered MCs. To better characterize the similarity data by the experts, in this work we introduced eight additional features which were demonstrated to have high discriminating power for cancer diagnosis [13]. Consequently, there were a total of 18 features used for describing the MCCs. For the purpose of selecting the most relevant features for similarity learning, we applied a feature selection procedure, called sequential backward selection [18]. The following set of 12 features was finally selected for characterizing a MC cluster:

1) compactness of the cluster: a measure of roundness of the region occupied by the cluster;
2) eccentricity of the cluster: the eccentricity of the smallest ellipse of the region (ratio of the distance between the foci and the major axis);
3) the number of MCs per unit area;
4) the average of the inter-distance between neighboring MCs.
5) the standard deviation of the inter-distance between neighboring MCs.
6) solidity of the cluster region: the ratio between cross-sectional area and the area of the convex hull formed by the MCs;
7) the moment signature of the cluster region: computed based on the distance deviation of the boundary point from the center of the region;
8) the number of MCs in the cluster;
9) the mean effective volume (area times effective thickness) of individual MCs;
10) the relative standard deviation of the effective thickness;
11) the relative standard deviation of the effective volume;
12) the second highest MC-shape-irregularity measure.

The details of these features can be found in [13]. In our experiment, all the feature components were normalized to have the same dynamic range (0,1).

III. Evaluation Study

To quantify the accuracy of the learned similarity function, we first computed the mean squared error (MSE) of the model compared to the observer scores on the 600 pairs scored in the observer study using a leave-one-out procedure. In addition, to evaluate the merit of the learning-based similarity measure for cancer diagnosis, we used the following two criteria: 1) cumulative neighbor matching rate and 2) multidimensional scaling (MDS). We compared the supervised learning-based similarity measure against two alternative distance measures. We describe these two criteria below in detail.

A. Cumulative Neighbor Matching Rate

We demonstrate the retrieval performance by using the so-called cumulative neighbor matching rate achieved by the learned similarity function as follows: for each query image, we compute the ratio of top k $(k = 1, 2, 3, 4 \ldots)$ images that actually match the disease condition of the query, and then average this ratio over all the queries (in our experiment each of the 200 images was used in turn as a query).

B. MDS as a Perceptual Evaluation Tool

MDS [17] is a powerful technique for representation and analysis of a set of objects based on their mutual similarity (or dissimilarity) measurements. The basic idea of MDS is to embed the objects of interest as points in a low-dimensional (typically 2- or 3-D) space such that the geometric distances between the points in this space are in accordance with the similarity measurements between the corresponding objects. In particular, Rubner [12] recently applied MDS as a visualization technique for retrieved images based on their texture and color distributions.

In order to evaluate the meaningfulness of our retrieval framework, we use MDS to embed the images in a 2-D space so that

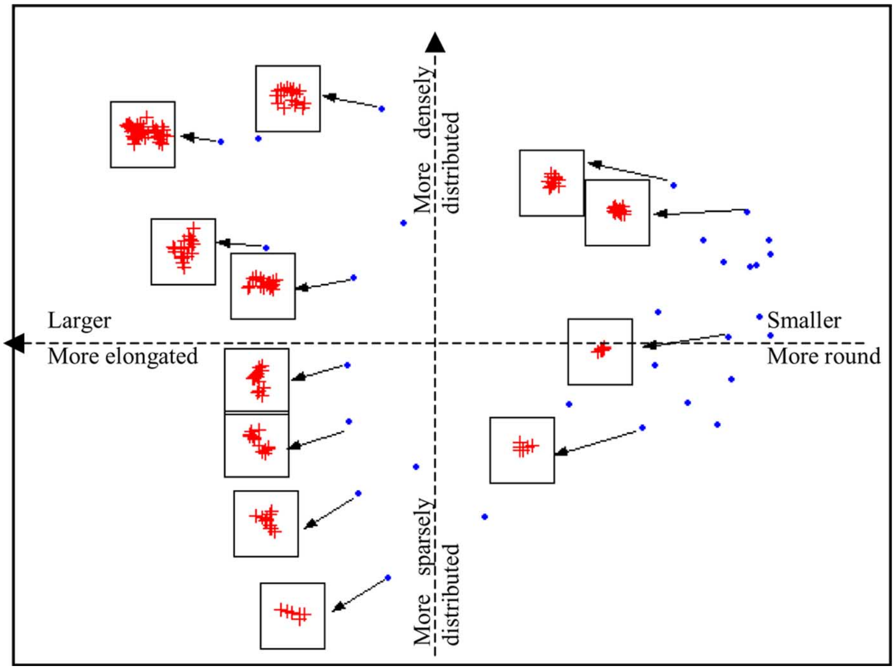

Fig. 7. MDS embedding of 30 MC clusters on a 2-D plane based on the observer similarity data.

distances in the embedding are as close as possible to the true distances between the images. In this way, it is convenient to visualize the retrieved mammograms within a local neighborhood of the query. In the MDS plot, both the query and retrieved mammograms will be displayed (as thumbnails) in a 2-D window according to their similarities. In this plot those mammograms most similar to the query will be placed most closely to the query while those less similar will be placed farther away. Besides being more intuitive, such an MDS approach will be much more informative, compared to a conventional approach that lists the retrieved images in the order of their similarities to the query. The MDS plot will not only show how similar the retrieved images are to the query, but more importantly, it will also reveal how the retrieved images are similar to each other.

In Fig. 7 we show an MDS plot of 30 regions of interest (ROIs) used in [9]. The solid dots represent the embedded locations of the ROIs in the 2-D plane; furthermore, higher similarity measures between the ROIs are inverse proportionally mapped to smaller distances between the corresponding points. As can be seen, the MDS plot reveals some rather interesting structure in the observer similarity data, as indicated by the two dashed lines. To assist the interpretation of the data, we have added these dashed lines to indicate how the ROIs are clustered in the plot according to the geometric distributions of the MC clusters (MCs are marked using cross).

In our retrieval framework, we will generate MDS plots according to learning-based similarity measure. Besides being a displaying tool for retrieved images, we can also apply MDS for browsing and exploring either all the images in a large database or only those images in a certain disease category. The MDS plot in such a case will produce a global view of these images, in which similar images will be clustered together according to certain image attributes. In a sense it could serve as a guide map for navigating and retrieving cases from the database.

C. Alternative Distance-Based Similarity Measures

We compare the supervised learning-based similarity measure against two alternative distance measures: 1) discriminant adaptive nearest neighbor (DANN) and 2) normalized cut (Ncut).

1) DANN Measure: DANN [19] is an improved version of the k nearest neighbor (KNN) measure based on the Euclidean distance [10], [20] for computing the similarity between two images. In DANN, a locally adaptive form of the nearest neighbor measure is used to ameliorate the curse of dimensionality. The distance between a point $\mathbf{x}'$ and a query $\mathbf{x}$ is defined as

$$D = (\mathbf{x}' - \mathbf{x})^T \Sigma (\mathbf{x}' - \mathbf{x}). \tag{8}$$

In KNN, the matrix Σ is the identity matrix I. In DANN, with a local discriminant model, the local within- and between-class covariance matrices are used to define the optimal shape of the neighborhood (Σ). In (8), the matrix Σ is computed as

$$\Sigma = W^{-1/2}[W^{-1/2}BW^{1/2} + \varepsilon I]W^{-1/2} \tag{9}$$

where B and W are the between- and within-class covariance matrices, respectively, which are computed from k_m local neighbors of $\mathbf{x}$. Specifically, $W = \sum_{k=1}^{2}\sum_{y_i=k}(\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T$ and $B = \sum_{k=1}^{2}(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})^T, \varepsilon$ is a small tuning parameter, and I is the identity matrix.

The matrix Σ is determined iteratively (initially with I) to obtain the adaptive shape of the neighborhood. It shrinks the neighborhood in directions the local class centroids differ, with the intention of ending up with a neighborhood in which the class centroids coincide. In DANN, k_m and ε are determined empirically from the data.

2) Ncut Measure: Normalized cut [21] originally is a method for image clustering. When applied for retrieval [22], it retrieves

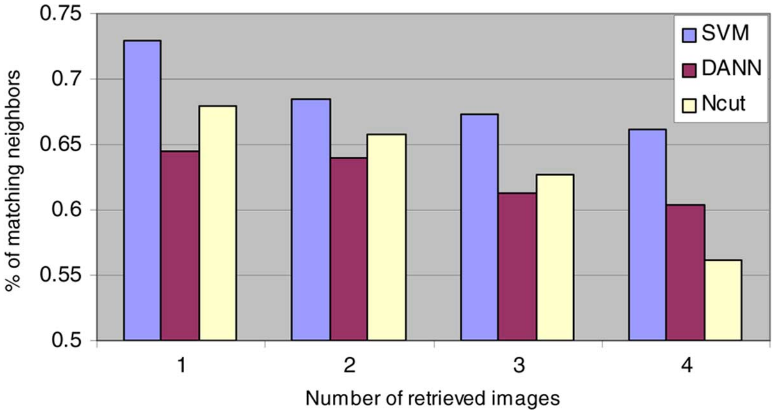

Fig. 8. Cumulative neighbor disease matching rates achieved by different retrieval measures.

a cluster of images rather than a set of ordered images. Here, we adapt this clustering method to mammogram retrieval.

In this method, a graph representation of the neighboring target images is defined as $G = (V, E)$, where the nodes $V = 1, 2, \ldots, n$ represent the images, and the edges $E = (i, j), i, j = 1, 2, \ldots, n$ are formed between every pair of nodes. Then non-negative weights w_{ij} are defined for the edges according to the similarity (a distance function; here the Euclidean distance is used as $w_{ij} = \exp(-d(i,j)^2)$) between the nodes. The weights are then organized into an affinity matrix. Clustering can then be formulated as a graph-partitioning problem. With this method, the images are organized into small groups so that the within-group similarity is high, and the between-group similarity is low. The images in the group where the query image resides are retrieved. Here we used a two-stage hierarchical fashion as in [22] to speed up the process for mammogram retrieval: in the first stage, the top k nearest neighboring images for a query image are treated as the target images; in the second stage, the Ncut clustering method is applied to the query image and its nearest neighboring target images to obtain the final retrieval results.

IV. Experiment Results and Discussions

The SVM similarity model described in Section II was trained using the observer data. Besides the human scores for the 600 image pairs, we also added the following pairs for training:

1) $\mathrm{SC}(\mathbf{u}, \mathbf{u}) = 10$;
2) $\mathrm{SC}(\mathbf{u}, \mathbf{v}) = 10$ if $\mathbf{u}$ and $\mathbf{v}$ are different views from the same case.

With a leave-one-out procedure, the SVM model achieved a MSE of 0.0334 per image pair compared to the observer scores.

Next, the trained SVM similarity model was tested with the 200 images in the database, where each of the 200 images in the dataset was used in turn as a query image. The average cumulative neighbor matching rate was calculated in the end. Similarly, we also applied the DANN and Ncut measures for retrieval. The test results are summarized in Fig. 8 for the different methods, where the average matching rate is plotted against k, which is the number of top retrieved images for each query. In particular, for SVM, the top most similar image can match the disease of the query 72.5% of the time. This is significantly different from the result by random pairing (of which the expected matching rate is 50.75%). As can be seen, the best performance was achieved by the SVM model. The parameter settings for the different methods were chosen by the leave-one-out retrieval procedure and shown as follows: SVM (Gaussian kernel, $\sigma = 1, C = 100, \varepsilon = 1$), DANN $(k_m = 40, \varepsilon = 1, k = 5)$, and Ncut $(k = 20)$.

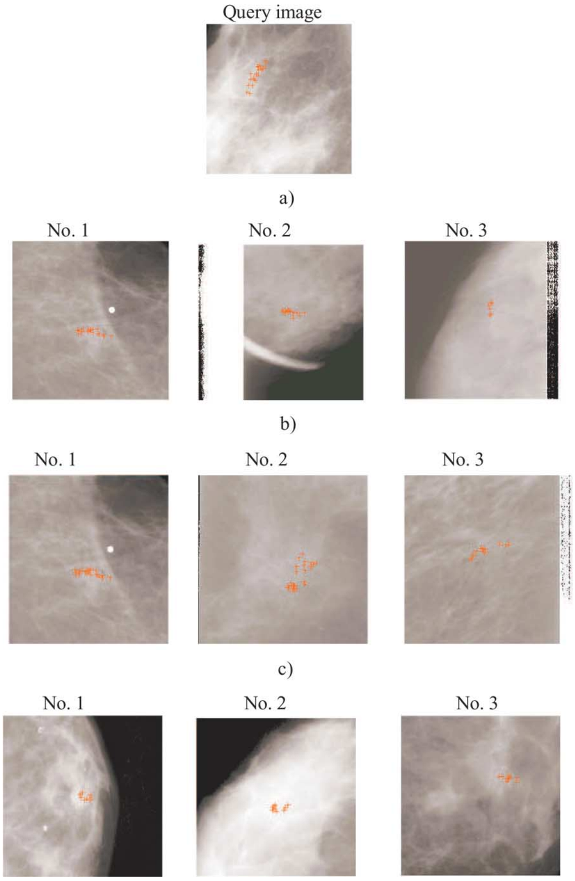

Fig. 9. Example 1: top three images retrieved by different similarity measure models. (a) SVM, (b) DANN, (c) Ncut.

In Figs. 9–11, we show some retrieval examples for three given query images by different similarity measure models. The individual MCs in each mammogram are marked out for

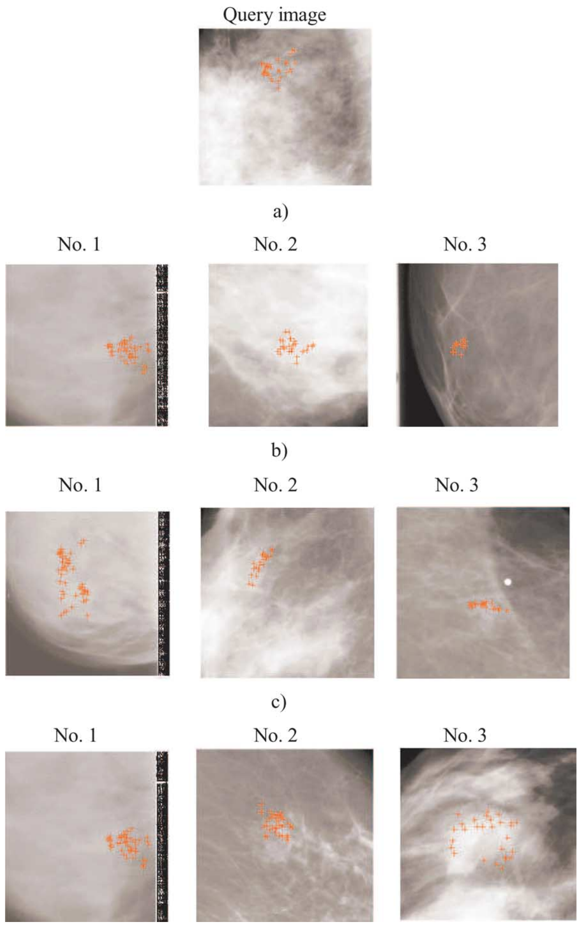

Fig. 10. Example 2: top three images retrieved by the different similarity measure models. (a) SVM, (b) DANN, (c) Ncut.

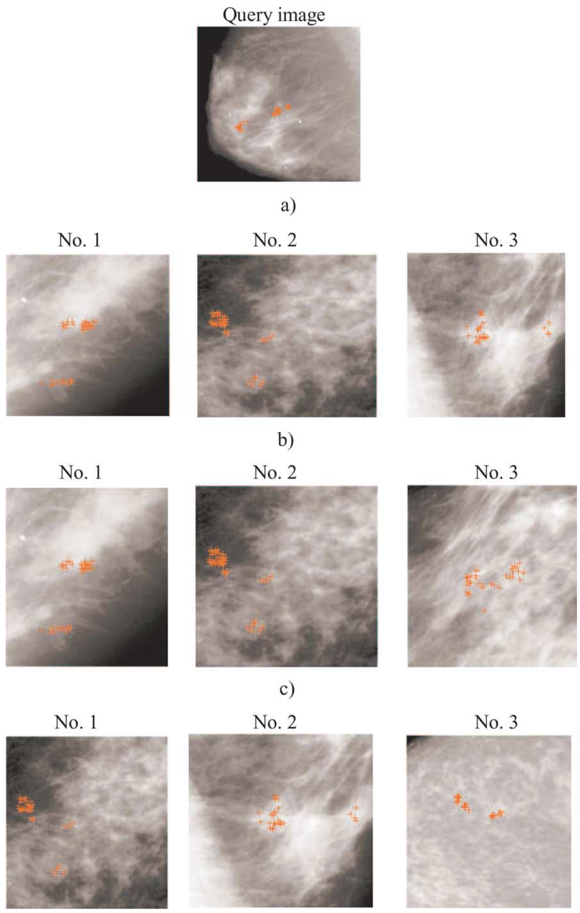

Fig. 11. Example 3: top three images retrieved by the different similarity measure models. (a) SVM, (b) DANN, (c) Ncut.

better visualization. As can be seen from these examples, the learning-based similarity measure can indeed achieve meaningful retrieval results and perform better than other similarity measure models. In constrast, DANN performed well in examples 1 and 3, and Ncut achieved good results in examples 2 and 3. Furthermore, in Figs. 12 and 13 we show MDS plots of retrieved results by the learning-based similarity measure for the first two examples.

V. Conclusion

In this work, we investigated a supervised learning approach for content-based mammogram retrieval based on expert observer similarity perception. The proposed similarity model was tested using a set of clinical mammograms. It was demonstrated to achieve significant improvement in retrieval performance over unsupervised learning methods. Encouraged by this success, in future work we plan to investigate whether such a system can serve as a more intuitive aid to radiologists.

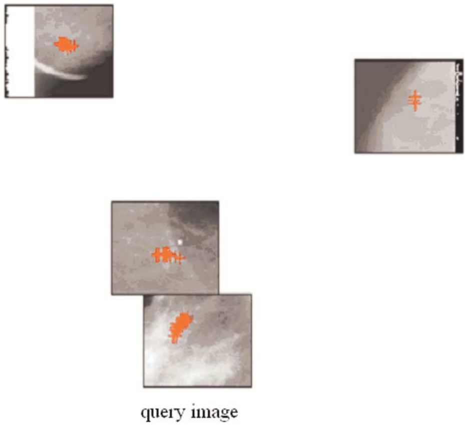

Fig. 12. Example 1: MDS display of retrieved results by learning-based similarity measure.

Appendix A
Symmetric SVM for Regression

To measure the similarity between two feature vectors, we want to learn a symmetric similarity function satisfying $f(\mathbf{u},\mathbf{v}) = f(\mathbf{v},\mathbf{u})$. For this purpose, we duplicate the image pairs, one with $(\mathbf{u},\mathbf{v})$ and the other with $(\mathbf{v},\mathbf{u})$. Then the cost function for SVM regression is

$$R(\mathbf{w},b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} L_\varepsilon(\mathbf{x}_i) + C\sum_{i=1}^{N} L_\varepsilon\left(\mathbf{x}_i^s\right). \quad (10)$$

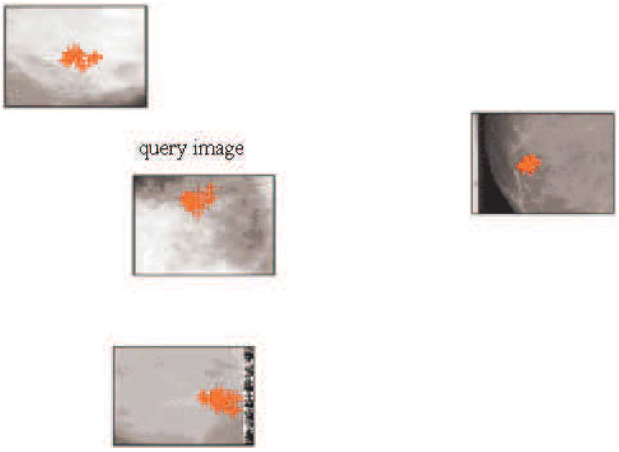

Fig. 13. Example 2: MDS display of retrieved results by learning-based similarity measure.

Here, $y(\mathbf{x}_i) = y(\mathbf{x}_i^s), \mathbf{x}_i = (\mathbf{u}_i^T, \mathbf{v}_i^T)^T, \mathbf{x}_i^s = (\mathbf{u}_i^T, \mathbf{v}_i^T)^T$. Accordingly, the Lagrangian function is defined as

$$\begin{aligned} &J\left(\mathbf{w}, \xi, \xi^{'}, \xi^s, \xi^{s'}, \alpha, \alpha^{'}, \alpha^s, \alpha^{s'}, \eta, \eta^{'}, \eta^s, \eta^{s'}\right) = \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &+ C\sum_{i=1}^{N}\left(\xi_i + \xi_i^{'}\right) + C\sum_{i=1}^{N}\left(\xi_i^s + \xi_i^{s'}\right) \\ &- \sum_{i=1}^{N}\left(\eta_i\xi_i + \eta_i^{'}\xi_i^{'}\right) - \sum_{i=1}^{N}\left(\eta_i^s\xi_i^s + \eta_i^{s'}\xi_i^{s'}\right) \\ &- \sum_{i=1}^{N}\alpha_i\left[\mathbf{w}^T\phi(\mathbf{x}_i) - y_i + \epsilon + \xi_i\right] \\ &- \sum_{i=1}^{N}\alpha_i^{'}\left[y_i - \mathbf{w}^T\phi(\mathbf{x}_i) + \epsilon + \xi_i^{'}\right] \\ &- \sum_{i=1}^{N}\alpha_i^s\left[\mathbf{w}^T\phi\left(\mathbf{x}_i^s\right) - y_i + \epsilon + \xi_i^s\right] \\ &- \sum_{i=1}^{N}\alpha_i^{s'}\left[y_i - \mathbf{w}^T\phi\left(\mathbf{x}_i^s\right) + \epsilon + \xi_i^{s'}\right]. \end{aligned} \tag{11}$$

The goal is to minimize J with respect to the weight vector $\mathbf{w}$ and slack variables $\xi, \xi^{'}$ and $\xi^s, \xi^{s'}$. By taking the partial derivatives and setting them to zero, we obtain

$$\mathbf{w} = \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^{'}\right)\phi(\mathbf{x}_i) + \sum_{i=1}^{N}\left(\alpha_i^s - \alpha_i^{s'}\right)\phi\left(\mathbf{x}_i^s\right) \tag{12}$$

$$\eta_i = C - \alpha_i \tag{13}$$

$$\eta_i^s = C - \alpha_i^s \tag{14}$$

$$\eta_i^{'} = C - \alpha_i^{'} \tag{15}$$

$$\eta_i^{s'} = C - \alpha_i^{s'}. \tag{16}$$

Next, substituting (12)–(16) into (11), we obtain the dual-optimization problem

$$\begin{aligned} Q\left(\alpha_i, \alpha_i^{'}, \alpha_i^s, \alpha_i^{s'}\right) &= \sum_{i=1}^{N} y_i\left[\left(\alpha_i - \alpha_i^{'}\right) + \left(\alpha_i^s - \alpha_i^{s'}\right)\right] \\ &- \epsilon\sum_{i=1}^{N}\left[\left(\alpha_i + \alpha_i^{'}\right) + \left(\alpha_i^s + \alpha_i^{s'}\right)\right] \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\alpha_i - \alpha_i^{'}\right)\left(\alpha_j - \alpha_j^{'}\right)K(\mathbf{x}_i, \mathbf{x}_j) \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\alpha_i^s - \alpha_i^{s'}\right)\left(\alpha_j^s - \alpha_j^{s'}\right)K\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right) \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\alpha_i - \alpha_i^{'}\right)\left(\alpha_j^s - \alpha_j^{s'}\right)K\left(\mathbf{x}_i, \mathbf{x}_j^s\right) \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\alpha_i^s - \alpha_i^{s'}\right)\left(\alpha_j - \alpha_j^{'}\right)K\left(\mathbf{x}_i^s, \mathbf{x}_j\right) \end{aligned} \tag{17}$$

$$\begin{aligned} \text{s.t.} &\sum_{i=1}^{N}\left(\alpha_i - \alpha_i^{'}\right) = 0 \\ &\sum_{i=1}^{N}\left(\alpha_i^s - \alpha_i^{s'}\right) = 0 \\ &0 \le \alpha_i \le C, i = 1, 2, \ldots, N \\ &0 \le \alpha_i^{'} \le C, i = 1, 2, \ldots, N \\ &0 \le \alpha_i^s \le C, i = 1, 2, \ldots, N \\ &0 \le \alpha_i^{s'} \le C, i = 1, 2, \ldots, N. \end{aligned}$$

Since $K(\mathbf{x}_i, \mathbf{x}_j^s) = K(\mathbf{x}_i^s, \mathbf{x}_j)$ and $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_i^s, \mathbf{x}_j^s)$, (17) is a symmetric function for $\alpha, \alpha\prime$ and $\alpha^s, \alpha^{s'}$. The final regression function can then be written as

$$f(\mathbf{x}) = \sum_{j=1}^{Ns}\gamma_j\left[K(\mathbf{x}_j, \mathbf{x}) + K\left(\mathbf{x}_j^s, \mathbf{x}\right)\right] + b \tag{18}$$

where $\gamma_j = \alpha_j - \alpha_j^{'} = \alpha_j^s - \alpha_j^{s'}$. Thus, if a training sample $\mathbf{x}_j$ is a support vector, then its symmetric sample $\mathbf{x}_j^s$ is also a support vector. Hence, $f(\mathbf{x}) = f(\mathbf{x}^s)$.

References

[1] A. W. M. Smeulders, M. Worrings, and S. Santini, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[2] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval system in medical applications-clinical benefits and future directions," *Int. J. Med. Informat.*, vol. 73, pp. 1–23, 2004.

[3] M. M. Rahman, T. Wang, and B. C. Desai, "Medical image retrieval and registration: Towards computer assisted diagnostic approach," in *Proc. IDEAS Workshop on Medical Information Systems: The Digital Hospital*, 2004, pp. 78–89.

[4] A. M. Knutzen and J. J. Gisvold, "Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions," *Mayo Clin. Proc.*, vol. 68, pp. 454–460, 1993.

[5] D. B. Kopans, "The positive predictive value of mammography," *Amer. J. Radiol.*, vol. 158, pp. 521–526, 1992.

[6] A. Swett and P. L. Miller, "Icon: A computer-based approach to differential diagnosis in radiology," *Radiology*, vol. 163, pp. 555–558, 1987.

[7] P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas, "Fast and effective retrieval of medical tumor shapes," *IEEE Trans. Knowl. Data Eng.*, vol. 10, no. 6, pp. 889–904, 1998.

[8] W. Guo, R. Chang, C. Lee, W. Moon, and D. Chen, "Retrieval technique for the diagnosis of solid breast tumors on sonogram," *Ultrasound in Med. and Biol.*, vol. 28, no. 7, pp. 903–909, 2002.

[9] I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, "A similarity learning approach to content based image retrieval: Application to digital mammography," *IEEE Trans. Med. Imag.*, vol. 23, no. 10, pp. 1233–1244, Oct. 2004.

[10] W. Niblack, R. Barber, and W. Equitz, "Querying images by content, using color, texture, and shape," in *Proc. SPIE*, 1993, vol. 1908, pp. 173–187.

[11] J. Hafner, H. S. Sawhney, and W. Equitz, "Efficient color histogram indexing for quadratic from distance functions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 7, pp. 729–736, Jul. 1995.

[12] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proc. IEEE Int. Conf. Computer Vision*, 1998, pp. 59–66.

[13] L. Wei, Y. Yang, and R. M. Nishikawa, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Trans. Med. Imag.*, vol. 24, no. 3, pp. 371–380, 2005.

[14] R. M. Nishikawa, Y. Yang, D. Huo, C. A. Sennett, J. Papaioannou, and L. Wei, "Observers' ability to judge the similarity of clustered microcalcifications on mammograms," in *Proc. SPIE*, 2004, vol. 5372, pp. 192–198.

[15] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[16] M. G. Kendall, *Rank Correlation Methods*, 4th ed. London, U.K.: Griffin, 1970.

[17] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer-Verlag, 1997.

[18] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. New York: Academic, 2003.

[19] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 6, pp. 607–616, Jun. 1996.

[20] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[21] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[22] Y. Chen, J. Z. Wang, and R. Krovetz, "Clue: Cluster-based retrieval of images by unsupervised learning," *IEEE Trans. Image Process.*, vol. 14, no. 8, pp. 1187–1201, Aug. 2005.

Liyang Wei (M'98) received the B.S. and M.S. degrees from the Department of Biomedical Engineering, Xi'an Jiaotong University, Xi'an, China, in 1998 and 2001, respectively, and the Ph.D. degree from the Department of Biomedical Engineering, Illinois Institute of Technology, Chicago, in 2006.

Her research interests include medical image analysis, machine learning, pattern recognition and computer-aided diagnosis.

Yongyi Yang (M'97–SM'03) received the B.S.E.E. and M.S.E.E. degrees from Northern Jiaotong University, Beijing, China, in 1985 and 1988, respectively, the M.S. degree in applied mathematics, and the Ph.D. degree in electrical engineering, both from Illinois Institute of Technology (IIT), Chicago, in 1992 and 1994, respectively.

He is currently on the faculty of the Department of Electrical and Computer Engineering at IIT, where he is a Professor. Previously, he was a faculty member with the Institute of Information Science, Northern Jiaotong University. His research interests are in signal and image processing, medical imaging, machine learning, pattern recognition, and biomedical applications. He is a co-author of *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics* (New York: Wiley, 1998).

Dr. Yang is an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Miles N. Wernick (M'92–SM'00) received the Ph.D. degree in optics from the University of Rochester, Rochester, NY, in 1990.

He was an NIH Postdoctoral Fellow in 1990 and Research Associate (Assistant Professor) from 1991 to 1994 in the Department of Radiology, University of Chicago, Chicago, IL. Since 1994, he has been with the Illinois Institute of Technology, Chicago, where he is a Professor of Electrical and Computer Engineering and Professor of Biomedical Engineering, and Director of the Medical Imaging Research Center. Since 2001, he has also been President of Predictek, Inc. His research interests are in medical imaging, image processing, and machine learning.

Dr. Wernick is currently Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING and the *SPIE/IS& T Journal of Electronic Imaging*, and a member of the IEEE BioImaging and Signal Processing (BISP) Technical Committee. He is co-editor of the book *Emission Tomography: The Fundamentals of PET and SPECT*.

Robert M. Nishikawa received the B.Sc. degree in physics in 1981 and the M.Sc. and Ph.D. degrees in medical biophysics in 1984 and 1990, respectively, all from the University of Toronto, Toronto, ON, Canada.

He is currently an Associate Professor in the Department of Radiology and is on the Committee on Medical Physics, both at the University of Chicago, Chicago, IL. He is also Director of the Carl J. Vyborny Translational Laboratory for Breast Imaging Research. His research interests are in computer-aided diagnosis, breast imaging, and evaluation of medical technologies. He is a member of the scientific advisory board for Dexela, Ltd., London, U.K.

Dr. Nishikawa is a Fellow of the American Association of Physicists in Medicine (AAPM).