A Similarity Learning Approach to Content-Based Image Retrieval: Application to Digital Mammography

Issam El-Naqa, *Member, IEEE*, Yongyi Yang*, *Senior Member, IEEE*, Nikolas P. Galatsanos, *Senior Member, IEEE*, Robert M. Nishikawa, and Miles N. Wernick, *Senior Member, IEEE*

***Abstract*—In this paper, we describe an approach to content-based retrieval of medical images from a database, and provide a preliminary demonstration of our approach as applied to retrieval of digital mammograms. Content-based image retrieval (CBIR) refers to the retrieval of images from a database using information derived from the images themselves, rather than solely from accompanying text indices. In the medical-imaging context, the ultimate aim of CBIR is to provide radiologists with a diagnostic aid in the form of a display of relevant past cases, along with proven pathology and other suitable information. CBIR may also be useful as a training tool for medical students and residents.**

The goal of information retrieval is to recall from a database information that is relevant to the user's query. The most challenging aspect of CBIR is the definition of relevance (similarity), which is used to guide the retrieval machine. In this paper, we pursue a new approach, in which similarity is learned from training examples provided by human observers. Specifically, we explore the use of neural networks and support vector machines to predict the user's notion of similarity. Within this framework we propose using a hierarchal learning approach, which consists of a cascade of a binary classifier and a regression module to optimize retrieval effectiveness and efficiency. We also explore how to incorporate online human interaction to achieve relevance feedback in this learning framework. Our experiments are based on a database consisting of 76 mammograms, all of which contain clustered microcalcifications (MCs). Our goal is to retrieve mammogram images containing similar MC clusters to that in a query. The performance of the retrieval system is evaluated using precision-recall curves computed using a cross-validation procedure. Our experimental results demonstrate that: 1) the learning framework can accurately predict the perceptual similarity reported by human observers, thereby serving as a basis for CBIR; 2) the learning-based framework can significantly outperform a simple distance-based similarity metric; 3) the use of the hierarchical two-stage network can improve retrieval performance; and 4) relevance feedback can be effectively incorporated into this learning framework to achieve improvement in retrieval precision based on online interaction with users; and 5) the retrieved images by the network can have predicting value for the disease condition of the query.

***Index Terms*—Computer-aided diagnosis, content-based image retrieval, digital radiography, kernel methods, mammogram, relevance feedback.**

I. Introduction

CONTENT-BASED IMAGE RETRIEVAL (CBIR) refers to the recall of images from a database that are relevant to a query, using information derived from the images themselves, rather than relying on accompanying text indices or other annotation. CBIR has received increasing attention as a result of the availability of large image databases in medicine, science, commerce, and the military [1], [2]. CBIR has been proposed to overcome the difficulties encountered in textual annotation for large image databases. Like a text-based search engine, a CBIR system aims to retrieve information that is relevant (or similar) to the user's query. In document retrieval, the query is usually a word or phrase; in CBIR, it is an image. The key to successful CBIR lies in the development of appropriate similarity metrics for ranking the relevance to the query image of images in a database. In CBIR, quantitative image features, computed automatically, are used to characterize image content. The image features may be extracted at either a low level (such as local edges [3]) or at a high level (such as a color histogram [4]), or both. The query image is then compared to the images in the database on the basis of the measured features. Those images in the database having the highest similarity to the query image are retrieved and displayed for the user.

The general application of image retrieval to broad image databases has experienced limited success, principally due to the difficulty of quantifying image similarity for unconstrained image classes (e.g., all images on the Internet). We expect that medical imaging will be an ideal application of CBIR, because of the more-limited definition of image classes (e.g., digital mammograms), and because the meaning and interpretation of medical images is better understood and characterized. In spite of this, the application of CBIR in medical imaging thus far has been somewhat limited [5]. In [6], a rule-based expert system was developed to display chest radiographs from a library of images as illustrative examples for helping radiologists' diagnosis. In [7], a retrieval method based on texture and shape

Manuscript received December 11, 2003; revised June 21, 2004. This work was supported by the National Institutes of Health (NIH)/National Cancer Institute (NCI) under Grant CA89668. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was H.-P. Chan. *Asterisk indicates corresponding author.*

I. El-Naqa is with the Medical School of Washington University, St. Louis, MO 63110 USA.

*Y. Yang is with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, 3301 South Dearborn Street, Chicago, IL 60616 USA (e-mail: yy@ece.iit.edu).

N. P. Galatsanos is with the Department of Computer Science, University of Ioannina, Ioannina, Greece.

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

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

Digital Object Identifier 10.1109/TMI.2004.834601

analysis was applied for search and retrieval of a database containing pulmonary computed tomography (CT) images. In [8], an algorithm was described for retrieval of three-dimensional (3-D) magnetic resonance images based on anatomical structure matching. In [9], a similarity metric based on Bayes decision theory was developed for retrieval of neuroradiological CT images. In [10] and [11], a technique was developed that reduces high-dimensional data to a two-dimensional feature space in which images that are close to each other are selected for purposes of visualizing relationships in the data. In [12], a retrieval method was developed using correlation coefficients in a database of pulmonary nodules represented by the joint histogram of the pattern CT density and 3-D curvature shape index.

A. A Learning Approach to Quantify Image Similarity

Unlike the existing approaches to CBIR, which are typically based on some simple distance measures for image similarity, we propose an approach in which machine-learning algorithms [neural networks and support vector machines (SVMs)] are trained to predict the measures of image similarity reported by human observers. We treat the learning of the similarity function as a nonlinear regression of the similarity coefficient on the features of the images. The method is developed using a set of 76 mammograms, all containing clustered microcalcifications (MCs). Our goal is to retrieve mammograms containing similar MC clusters to that in a query mammogram. The proposed retrieval framework is evaluated statistically using a cross-validation procedure.

The feasibility of a learning-based approach for modeling perceptual similarity was first demonstrated in our previous work in [13] using simulated image data. In this paper, we expand this approach in two major aspects. First, we develop a hierarchal two-stage learning network for improved performance. Second, we explore how to utilize user interaction, known as relevance feedback, in the learning framework so as to achieve online adaptation to the user.

B. Application to Mammography

Mammography has been by far the most effective means for early detection of breast cancer, a leading cause of death in women in many developed countries. The sensitivity of mammography is approximately 90% [14]. In spite of the technological advances in recent years, mammogram reading still remains a difficult clinical task. Some breast cancers may produce changes in mammograms that are subtle and difficult to recognize. It has been reported that 10%–30% of lesions are misinterpreted during routine screening of mammograms [15]. Furthermore, it is very difficult to distinguish benign lesions from malignant ones in mammograms. As a result, between 2 and 10 women are biopsied for every cancer detected, causing needless fear and pain to women who are biopsied [16], [17]. This low specificity results in a relatively large interobserver variability that can lead to failure to biopsy malignant lesions and potentially avoidable biopsy of benign lesions [18].

We conjecture that by presenting images with known pathology that are "visually similar" to the image being evaluated, the use of a mammogram retrieval system may provide a more intuitive aid to radiologists, potentially leading to improvement in their diagnostic accuracy. Furthermore, it is expected that the proposed technique would be a useful aid in the training of students and residents, since it would allow them to view images of lesions that appear similar, but may have differing pathology.

An alternative approach to computer-aided diagnosis (CAD), in which the likelihood of malignancy is computed (e.g., [19]), has been studied to a large extent in the literature. The proposed retrieval system is in principle very different, and may helpfully complement existing diagnostic aids. Our retrieval system follows a "critiquing" approach [20]: instead of proposing a diagnosis, it aims to assist the radiologist by providing relevant supporting evidence from prior known cases. If we view the human observer as a classifier, then the aim of the CBIR system is to provide the observer with training-set examples that are close to his decision boundary, along with the correct class labels (proven pathology) for these examples. The hypothesis that we ultimately hope to demonstrate is that this approach will improve the classification (diagnostic) performance of the observer.

In developing a CBIR system for digital mammography, we argue that the similarity metric must conform closely to the user's notions of similarity, and that simple, mathematical distance metrics may not be adequate for describing perceived similarity. Therefore, we aim to show that a learned concept of similarity can outperform simple distance metrics in modeling the user's similarity concept.

The remainder of the paper is organized as follows. First, an overview of the proposed learning-based retrieval framework is provided in Section II. In Section III, the hierarchical learning network is described, and relevance-feedback techniques are developed in Section IV. An evaluation study, including data-set acquisition, training, and testing procedures is described in Section V. Experimental results are presented in Section VI. Finally, conclusions are drawn in Section VII.

II. Overview of the Proposed Image-Retrieval Framework

We assume that the user's notion of similarity between a pair of images is a function of the relevant features in the images. We then use machine learning to model this notion of similarity for the purpose of CBIR. Our goal is to find, among the many images in the database, those that are visually most similar to the query as judged by the user.

The proposed framework is illustrated with a functional diagram in Fig. 1. For a given query image, we first characterize its content by an M-dimensional vector $\mathbf{u}$, quantifying the key relevant features of the image. This feature vector is then compared to the corresponding feature vector $\mathbf{v}$ of a database entry by way of a learning machine, denoted by a nonlinear mapping $f(\mathbf{u}, \mathbf{v})$, to produce a similarity coefficient (SC). The images with the highest SCs (say, those above a prescribed threshold value T) are then retrieved from the database.

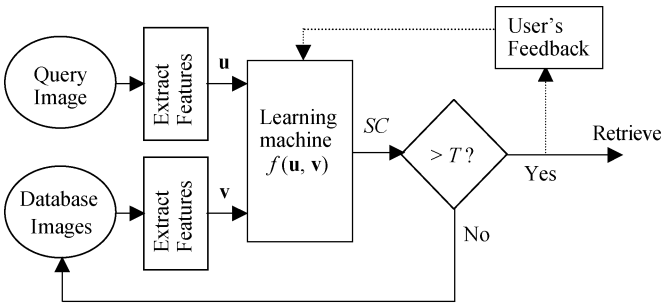

Fig. 1. Proposed image retrieval framework with relevance feedback.

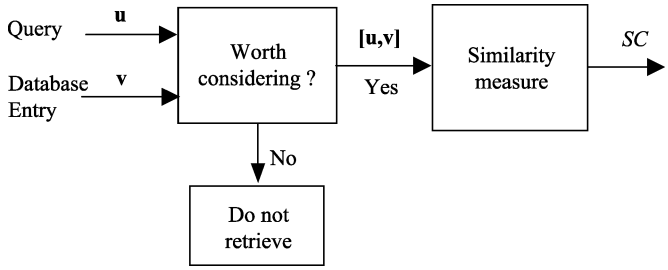

Fig. 2. The two-stage hierarchical learnig framework.

Clearly, the key to this framework lies in the learning machine $f(\mathbf{u},\mathbf{v})$. Of course, an equally important issue, if not more important, is the selection of feature vector $\mathbf{u}$ so that its components are relevant to perceptual similarity. In this paper, we will make use of existing features already in use for CAD in the literature (Section V-C).

Ideally, the learning machine should have the following properties: 1) $f(\mathbf{u},\mathbf{v})$ must closely conform to the user's notion of similarity; 2) the learning machine should involve reasonable computational complexity so that it can be applied to a large-scale database; and 3) $f(\mathbf{u},\mathbf{v})$ should provide the user with the ability to refine the search in a process called relevance feedback (indicated by the dashed path in Fig. 1).

We adopt a supervised learning approach for $f(\mathbf{u},\mathbf{v})$. For this purpose we first collect a set of sample image pairs, each having a labeled SC (e.g., obtained from observer studies). We then train a learning machine $f(\mathbf{u},\mathbf{v})$ with these samples. Specifically, letting $SC(\mathbf{u},\mathbf{v})$ denote the similarity coefficient between an image pair $\mathbf{u}$ and $\mathbf{v}$, we model $SC(\mathbf{u},\mathbf{v})$ as

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

where ξ is the modeling error of the learning machine. Our aim is to determine a functional $f(\cdot,\cdot)$ that will generalize well to images outside the training set.

Since our aim is always the comparison of pairs of images, we will view the similarity metric as a functional of a single argument $\mathbf{x} \equiv \binom{\mathbf{u}}{\mathbf{v}}$, which is a concatenation of the feature vectors $\mathbf{u}$ and $\mathbf{v}$ of two images to be compared; thus, we redefine the similarity functional $f(\mathbf{u},\mathbf{v})$ as $f(\mathbf{x})$.

III. Hierarchical Learning Network

We propose to use a two-stage hierarchical learning network to model the perceptual similarity for retrieval. This network consists of a cascade of a binary classifier stage and a regression stage for predicting the SCs between a query image and the images in the database, as illustrated in Fig. 2. In the first stage, images that are very different from the query image are eliminated from further consideration by a binary classifier. Images surviving this stage are then compared to the query in the second stage to obtain a numerical SC for retrieval.

The learning network in Fig. 2 is hierarchical in the following sense: during the training phase, the first-stage classifier functions as a coarse, binary learner, the purpose of which is for triage; i.e., the first stage identifies simply whether a database entry is sufficiently similar to the query for further consideration. The second stage functions as a more-refined learner, the purpose of which is to measure quantitatively the similarity between a surviving database entry and the query.

The reasons for this approach are as follows. First, the triage classifier avoids the computational cost of carefully measuring the SCs of those database entries that are not at all similar to the query, and thus will certainly not be determined to be relevant. Second, by training the second stage using only reasonably similar pairs, the learning machine can be better fine-tuned to predict SCs for those image pairs that are of genuine interest.

Of course, the use of a triage stage can also have adverse effect, i.e., it may eliminate some truly similar images from further consideration. To ameliorate this effect, in the following we will modify the cost function of the SVM classifier such that it will impose a greater penalty on missed similar images than on misclassified nonsimilar images. As demonstrated by our experimental results (Section VI), this approach can lead to significant improvement in retrieval performance. Below we discuss the details of the two-stage network.

A. First-Stage Classifier

Consider a query image and a database entry with feature vectors $\mathbf{u}$ and $\mathbf{v}$, respectively. The task of the first-stage classifier is to determine whether the two images are sufficiently similar for further consideration. This is treated as a two-class pattern classification problem, i.e., the mammogram image pair is either reasonably similar (designated as "class 1") or not similar (designated as "class 2"). For reasons of computational speed, we employ a linear classifier for this task, i.e., we use a decision function of the form

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

such that $h(\mathbf{x}) \geq 0$ if the image pair $\mathbf{x}$ is sufficiently similar, and $h(\mathbf{x}) < 0$ otherwise. In other words, image pairs from the two different classes are separated by the hyperplane $h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0$.

The decision function $h(\mathbf{x})$ is to be determined from training samples. Let $\{(\mathbf{x}_i, d_i), i = 1,2,\ldots,l\}$ denote a given set of l training examples, where each sample pair $\mathbf{x}_i$ has a known class label d_i (i.e., $d_i = +1$ for class 1, and $d_i = -1$ for class 2). The problem then is how to determine $\mathbf{w}$ and b in $h(\mathbf{x})$ so that it that can correctly classify an input pattern (not necessarily from the training set).

We consider two types of pattern classifiers for this task: 1) a Fisher discriminant and 2) an SVM. To distinguish between the two classifiers, below we use $\mathbf{w}_1$ and b_1 to denote the parameters $\mathbf{w}$ and b in $h(\mathbf{x})$ for the Fisher discriminant, and use $\mathbf{w}$ and b for the SVM.

1) Fisher Discriminant Classifier: The Fisher discriminant is based on the principle of projecting the data onto a one-dimensional space so that the two classes are well separated [21]. The discriminant vector $\mathbf{w}_1$ in the decision function is determined by

$$\mathbf{w}_1 = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}) \tag{3}$$

where $\boldsymbol{\mu}_{+1}$ and $\boldsymbol{\mu}_{-1}$ are the mean vectors of the two classes, and $\mathbf{\Sigma}$ is the total within-class covariance matrix, all estimated from the training samples. The constant b_1 is computed as

$$b_1 = -\frac{1}{2}\mathbf{w}_1^T(\boldsymbol{\mu}_{+1} + \boldsymbol{\mu}_{-1}). \tag{4}$$

2) SVM Classifier: SVM is a constructive learning procedure based on statistical learning theory [22]. It is based on the principle of structural risk minimization, which aims at minimizing the bound on the generalization error (i.e., error made by the learning machine on data unseen during training) rather than minimizing the mean square error over the data set. As a result, an SVM tends to perform well when applied to data outside the training set. In recent years, SVM learning has found a wide range of real-world applications (see, for example, [23]–[27]). In many of these applications it has been reported that SVM-based approaches are able to outperform competing methods. In our own work [28], we developed an SVM-based approach for detection of microcalcifications in mammograms, and demonstrated using clinical mammogram data that such an approach could outperform several well-known methods in the literature.

Using the training data set $\{(\mathbf{x}_i, d_i), i = 1, 2, \ldots, l\}$, a linear SVM classifier in its original form is formulated as minimization of the following cost function:

$$\begin{aligned} J(\mathbf{w}, \boldsymbol{\xi}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to } & d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0; \quad i = 1, 2, \ldots, l \end{aligned} \tag{5}$$

where C is a user-specified, positive parameter, and ξ_i are slack variables.

The cost function in (5) constitutes the so-called *structural risk*. It consists of both the empirical risk (i.e., the training errors reflected by the second term) and the model complexity measure (the first term). The regularization parameter C in (5) is used to define the tradeoff between these two factors. In particular, when the two classes are separable, the SVM classifier amounts to maximize the separating margin between the two classes [as illustrated in Fig. 3(a)].

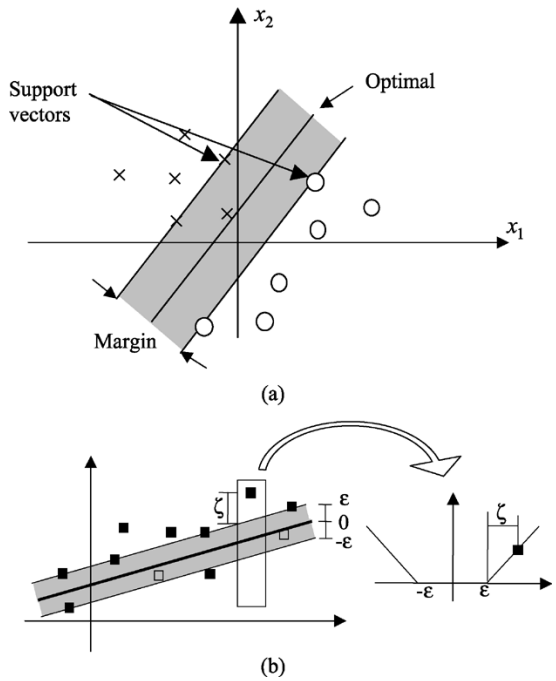

Fig. 3. Illustration of SVMs: (a) classification with a linear hyperplane that maximizes the margin between the two classes; and (b) ε-insensitive SVM for regression, where the loss function does not penalize errors below the parameter ε. The support vectors are indicated by filled squares.

For our task at hand, we propose to modify the SVM cost function in (5) as

$$J(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C^+\sum_{i\in Z^+}\xi_i + C^-\sum_{i\in Z^-}\xi_i, \tag{6}$$

where $C^+ > C^-$, and Z^+, Z^- are the index sets of the training samples belonging to class 1 (i.e., $d_i = +1$) and class 2 (i.e, $d_i = -1$), respectively. This imposes a greater penalty (C^+) on missed similar images than on misclassified nonsimilar images (C^-). The rationale is that the first-stage classifier is for prescreening only and should be designed to pass marginal cases to the second stage for further consideration.

Using the technique of Lagrange multipliers, one can show that a necessary condition for minimizing $J(\mathbf{w}, \xi)$ in (6) is that the vector $\mathbf{w}$ is formed by a linear combination of the vectors $\mathbf{x}_i$, i.e.,

$$\mathbf{w} = \sum_{i=1}^{l}\alpha_i d_i \mathbf{x}_i \tag{7}$$

where $\alpha_i \geq 0$, $i = 1, 2, \ldots, l$ are the Lagrange multipliers associated with the constraints in (5).

The Lagrange multipliers $\alpha_i \geq 0$, $i = 1, 2, \ldots, l$, are solved from the dual form of (6), which is expressed as

$$\max \tilde{J}(\alpha_1, \alpha_2, \ldots, \alpha_l) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j d_i d_j \mathbf{x}_i^T\mathbf{x}_j \tag{8}$$

subject to the constraints

$$1)\ \sum_{i=1}^{l} \alpha_i d_i = 0$$
$$2)\ 0 \le \alpha_i \le C^+, \quad \text{for } d_i = +1$$

or

$$0 \le \alpha_i \le C^-, \quad \text{for } d_i = -1, \quad i = 1, 2, \ldots, l. \quad (9)$$

The cost function $\tilde{J}(\alpha_1, \alpha_2, \ldots, \alpha_1)$ is convex and quadratic in terms of the unknown parameters α_i. In practice, the maximization in (8) is solved numerically through quadratic programming [22].

Analytic solutions of (8) are not readily available, but it is still informative to examine the conditions under which an optimal solution is achieved. The Karush–Kuhn–Tucker (KKT) optimality conditions for (8) lead to the following three cases for each α_i.

1) $\alpha_i = 0$. This corresponds to $d_i(\mathbf{w}^T\mathbf{x}_i + b) > 1$. In this case, the data element $\mathbf{x}_i$ is outside the decision margin of the function $h(\mathbf{x})$ and is correctly classified.
2) $0 < \alpha_i < C^+$ for $d_i = +1$; or $0 < \alpha_i < C^-$ for $d_i = -1$. In this case, $\mathbf{w}^T\mathbf{x}_i + b = d_i$. The data element $\mathbf{x}_i$ is strictly located on the decision margin of $h(\mathbf{x})$. Hence, $\mathbf{x}_i$ is called a *margin support vector* of $h(\mathbf{x})$.
3) $\alpha_i = C^+$ for $d_i = +1$; or $\alpha_i = C^-$ for $d_i = -1$. In this case, $d_i(\mathbf{w}^T\mathbf{x}_i + b) < 1$. The data element $\mathbf{x}_i$ is inside the decision margin (though it may still be correctly classified). Accordingly, $\mathbf{x}_i$ is called an *error support vector* of $h(\mathbf{x})$.

It is typical that most of the training examples are correctly classified by the trained classifier (case 1), i.e., only a few training examples will be support vectors. For simplicity, let $\mathbf{s}_j$, α_j^*, $j = 1, 2, \ldots, l_s$, denote these support vectors and their corresponding nonzero Lagrange multipliers, respectively, and let d_j denote their class labels. The SVM decision function can thus be simplified as

$$h(\mathbf{x}) = \sum_{j=1}^{l_s} \left(\alpha_j^* d_j\right) \mathbf{s}_j^T \mathbf{x} + b. \quad (10)$$

Note that the decision function is now determined directly by support vectors $\mathbf{s}_j$, $j = 1, 2, \ldots, l_s$, which are determined by solving the optimization problem in (8) during the training phase.

B. Regression Stage

The regression stage is used to provide quantitative SCs between the query and those images deemed sufficiently similar by the classification stage. Consequently, only a subset of the training data will be qualified for the training of the learning machine in this stage. In this paper, we consider the following two approaches for learning the similarity function $f(\mathbf{x})$: 1) an SVM and 2) a general regression neural network (GRNN) [29].

1) SVM Regression: SVM learning can also be applied for regression. An SVM formulation in such a case maintains many of the characteristics of the classification case. For nonlinear regression, an SVM in concept first maps the input data vector $\mathbf{x}$ into a higher dimensional space $\mathbf{H}$ through an underlying nonlinear mapping $\Phi(\cdot)$; then applies a linear regression in this mapped space. That is, a nonlinear SVM regression function can be written in the following form:

$$f(\mathbf{x}) = \mathbf{w}^T\Phi(\mathbf{x}) + b. \quad (11)$$

Let $\{(\mathbf{x}_i, y_i), i = 1, 2, \ldots, l'\}$ denote a set of training samples surviving the first stage, where y_i is the human-observer SC for the image pair denoted by $\mathbf{x}_i$. The parameters $\mathbf{w}$ and b in the regression function in (11) are determined through minimization of the following structural risk:

$$R(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l'} L_\varepsilon(\mathbf{x}_i) \quad (12)$$

where $L_\varepsilon(\cdot)$ is the so-called ε-*insensitive loss function* which is defined as

$$L_\varepsilon(\mathbf{x}) = \begin{cases} |y - f(\mathbf{x})| - \varepsilon, & \text{if } |y - f(\mathbf{x})| \ge \varepsilon \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

The function $L_\varepsilon(\cdot)$ has the property that it does not penalize errors below the parameter ε, as illustrated in Fig. 3(b). The constant C in (12) determines the tradeoff between the model complexity and the training error.

As with the case of classification, the regression function $f(\mathbf{x})$ in (11) is also characterized by the support vectors. It can be written as follows:

$$f(\mathbf{x}) = \sum_{i=1}^{l'_s} \gamma_i K(\mathbf{x}, \mathbf{s}_i) + b \quad (14)$$

where $\mathbf{s}_i$, $i = 1, 2 \cdots l'_s$, denote the support vectors, and $K(\mathbf{x}, \mathbf{s}_i) \equiv \Phi(\mathbf{x})^T\Phi(\mathbf{s}_i)$ which is called a kernel function. A training sample $(\mathbf{x}_i, y_i)$ is a *margin support vector* when $|f(\mathbf{x}_i) - y_i| = \varepsilon$, and an *error support vector* when $|f(\mathbf{x}_i) - y_i| > \varepsilon$.

From (14), we can directly evaluate the regression function through the kernel function $K(\cdot, \cdot)$ without the need to specifically addressing the underlying mapping $\Phi(\cdot)$. In this paper, we consider two kernel types: polynomial kernels and Gaussian radial basis functions (RBF). These are among the most commonly used kernels in SVM research, and are known to satisfy Mercer's condition [22]. They are defined as follows.

1) Polynomial kernel

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^p \quad (15)$$

where $p > 0$ is a constant that defines the kernel order.

2) RBF kernel

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) \quad (16)$$

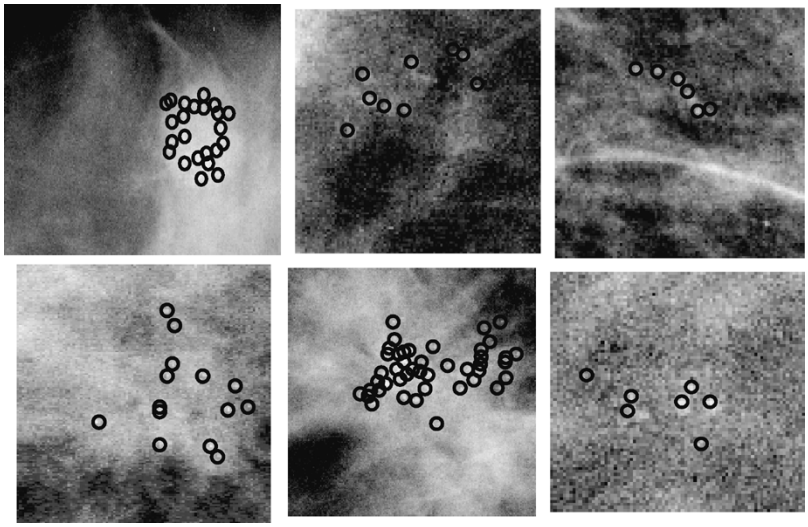

Fig. 4. Examples of mammogram regions containing clustered microcalcifications (indicated by circles).

where $\sigma > 0$ is a constant that defines the kernel width.

2) GRNN Regression: The GRNN computes an estimate of the conditional mean of the SC for an image pair from the human-observer data [29]. It is based on an estimate of the joint probability density of the input and the output obtained by the Parzen method [29]. With training data $\{(\mathbf{x}_i, y_i), i = 1, 2, \ldots, l'\}$, the output of the GRNN can be represented as

$$f(\mathbf{x}) = \frac{\sum_{i=1}^{l'} y_i \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma^2}\right)}{\sum_{i=1}^{l'} \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma^2}\right)} \tag{17}$$

where $\sigma > 0$ defines the kernel width.

Note that the GRNN estimate $f(\mathbf{x})$ in (17) has a similar form to the SVM estimate in (14) when the RBF kernel is used. The major difference between the two, however, is that only the support vector samples are used in the SVM in (14), while all the training samples are used in the GRNN in (17). Thus, the SVM estimate can be computationally advantageous over the GRNN.

IV. Relevance Feedback

In this section, we explore how to incorporate relevance feedback into our proposed learning-based retrieval approach. Relevance feedback is a post-query process to refine the search by using positive and/or negative indications from the user of the relevance of retrieved images. It has been applied successfully in traditional text-retrieval systems for improving the results of a retrieval strategy [30]. In particular, we consider the following scenario: for a query image $\mathbf{q}$ a user selects a relevant image $\mathbf{r}$ amongst the retrieved images to confirm that the retrieved $\mathbf{r}$ is indeed similar to the query $\mathbf{q}$; we want to incorporate this information to further refine the search, hoping that more relevant images could be found for the same query $\mathbf{q}$.

In this paper, we consider the following simple approach for relevance feedback: we explicitly incorporate the impact of the feedback image $\mathbf{r}$ in the measure of similarity between the query image $\mathbf{q}$ and a database entry $\mathbf{d}$. Specifically, we use the following weighted SC:

$$\widetilde{SC}(\mathbf{q}, \mathbf{d}) = (1 - w) \cdot SC(\mathbf{q}, \mathbf{d}) + w \cdot SC(\mathbf{r}, \mathbf{d}) \tag{18}$$

where w is a weighting parameter used to adjust the relative impact of the feedback image $\mathbf{r}$. The images with the highest weighted SCs are then retrieved.

An alternative to the above weighting approach is to adapt the learning machine based on the feedback information. We will consider this in a separate study [31], as the main goal of this paper is to demonstrate the feasibility of a learning framework for similarity modeling.

V. Performance Evaluation Study

A. Mammogram Data Set

The proposed retrieval framework was developed and tested using a data set collected by the Department of Radiology at The University of Chicago. This data set consists of 76 clinical mammograms, all containing multiple MCs. These mammograms are of dimension 1000 × 700 pixels, with a spatial resolution of 0.1 mm/pixel and 10-bit grayscale. Collectively, there are a total of 1120 MCs in these mammograms, which were identified by a group of experienced mammographers.

MCs are tiny calcium deposits that appear as small bright spots (typically 0.05–1 mm in diameter) in a mammogram. MC clusters (MCCs) in a mammogram provide valuable information to radiologists in diagnosis of cancer. For example, linearly distributed MCCs are typically malignant, while round clusters are typically benign [32]. In Fig. 4, we show a number of different regions of interest (ROIs) extracted from the mammograms in the data set, all of which contain MCCs.

Our objective is to apply the proposed framework to retrieve mammograms containing similar MCCs to that in a query mammogram.

B. Observer Similarity Data

For the training and testing of the algorithms, ROIs containing the identified MCCs were first extracted from all the mammograms in the data set (as shown in Fig. 4). Among the 76 mammograms, 74 contain only a single ROI, while the other two have two ROIs. These MCC ROIs were then used in a subsequent observer study to obtain SCs for different ROI pairs, which were then used to form training and testing samples.

The observer study was carried out by a panel of six human observers, who scored the similarity between each pair of ROIs based on their geometric distributions on a scale from 0 (most dissimilar) to 10 (most similar). It consisted of the following different sessions: 1) a "precalibration" session; 2) individual scoring sessions; and 3) a statistical analysis session for both intraobserver and interobserver consistencies.

The panel of observers first participated in a "precalibration" session ($\sim$1 h), the goal of which was to establish a consensus among the observers on a uniform measure of the perceptual similarity and to identify tentative "anchor pairs" (prototype examples) along the scale (from very different to very similar, all chosen randomly from the mammogram set).

For the individual scoring sessions, we randomly selected 30 ROIs from the mammogram set, each of which corresponds to a different patient. The observers then scored the similarity for all the possible pairs (a total of 435) formed by these ROIs, assisted by a software user interface. In each session, a query ROI was displayed along with up to 15 other ROIs simultaneously on the same computer screen (presented in a random order). The observer then assigned a continuous SC value between the query and each of the other ROIs by mouse-clicking on a thermometer bar on the computer screen. Each of the 30 ROIs in the dataset was used in turn as the query, yielding a total of 870 SCs from each observer (each MCC pair was scored twice).

The collected SCs were then analyzed for both intraobserver and interobserver consistencies. Specifically, the Kendall's rank correlation method [33] was first applied to test the consistency between the two scores reported by each observer for each of the 435 MCC pairs. The two scores were then averaged for each pair. Afterward, the Kendall's rank correlation method was applied to analyze the interobserver consistency among the six observers. The scores reported by the six observers were then averaged for each of the 435 pairs to obtain the SCs.

Specifically, for intraobserver consistency, Spearman's rank correlation statistic ρ was computed to be 0.7551, 0.7241, 0.6675, 0.7156, 0.7827, and 0.6850, respectively for the six observers; we also computed, using Fisher's transformation [34], the corresponding 95% confidence intervals of these coefficients to be [0.7102, 0.7938], [0.6746, 0.7671], [0.6101, 0.7179], [0.6648, 0.7598], [0.7421, 0.8176], and [0.6300, 0.7332], respectively. For the interobserver consistency, Kendall's *coefficient of concordance* W was computed to be 0.25. The coefficient W was computed by tabulating together all the ranking scores from all the six observers. As explained by Kendall [33], W is a measure of "the communality of the judgments for the m (6 in our case) observers". Specifically, W is calculated in the following two steps: 1) for each MCC pair, the sum of the ranking scores given by all the observers is computed, and the deviation by this sum from its mean value (assuming completely independent judgments by the observers) is computed; 2) the sum of squares of all these deviations is then computed (and adjusted by a constant factor) to yield W. A test of significance (the null hypothesis being that all observers are independent in their judgments) indicates that this result is statistically significant with its p-value below 0.0001; that is, under the null hypothesis, the probability of obtaining a value as great as or greater than 0.25 for W is less than 0.0001.

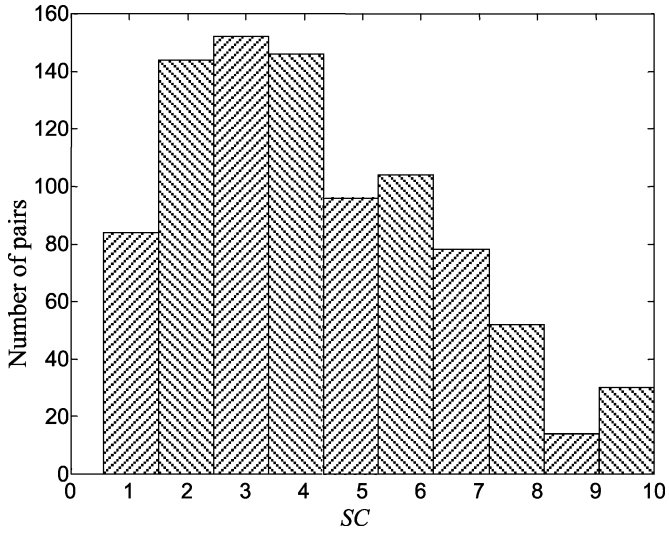

Fig. 5. Histogram plot of the observers' similarity scores (SCs) for the 465 MCC pairs in the dataset.

Finally, we introduced 30 "ideal" pairs, formed by each query ROI with itself. These pairs were all assigned a perfect SC (10).

In summary, a total of 465 MCC pairs were scored and recorded. In Fig. 5, we show a histogram plot of the obtained observers' similarity scores for the 465 pairs.

All the six observers have backgrounds in medical image analysis, with one of them also having a background in medical physics. To facilitate the observer study, all the individual MCs were clearly marked out in all the ROIs involved (based on the experts' readings). While there are other potentially important image features one might consider, we elected in this preliminary demonstration to retrieve images based on the spatial characteristics of the clusters alone. This will enable us to demonstrate the feasibility of the proposed framework using an observer-data set of reasonable size.

C. Extraction of MCC Features

To describe the geometric features of MCCs, we started with a total of 25 shape descriptors, most of which were used in the literature for shape analysis of MCCs [35]–[37]. We then applied a so-called *sequential backward selection* procedure [38] to reduce the set of salient features down to nine, which we describe below in detail. We point out that features 1, 4, and 5 were used in [35], [36], and features 8 and 9 were used in [37]. Furthermore, we note that the above feature-selection process was performed using an observer-dataset obtained using simulated distributions of clustered microcalcifications [13], used in our early development of the learning-based similarity framework. As demonstrated in this paper, these features can also lead to

good predictability in the observer SCs using clinical mammograms.

1) *Cross sectional area* (A): the area occupied by the cluster. It is computed in the following steps: 1) a binary image is first created in which the pixels corresponding to the centers of the MCs are set to 1 and all the rest of the pixels set to 0. 2) a Delaunay triangulation is next applied to connect the centers of the MCs in this binary image; the average interdistance between neighboring MCs, denoted by $\overline{p}$, is then computed based on this triangulation. 3) a morphological *closing* operation with a circular structuring element having a radius of $\overline{p}$ is then performed on the binary image to fill the gaps among the MCs. The area of the resulting region is then computed.
2) *Compactness*: a measure of roundness of the region occupied by the cluster. It is computed as
$$C_f = \frac{4\pi A_f}{P_f^2} \tag{19}$$
A_f, P_f are the area and perimeter of the solid region occupied by the cluster (i.e., holes are filled when necessary), respectively. Note that A_f will differ from the cross sectional area A when the occupied region contains any holes.
3) *Eccentricity*: the eccentricity of the smallest enclosing ellipse of the region, computed as the ratio of the distance between the foci and the length of the major axis of the ellipse.
4) *Density*: the spatial density of the MCs in the cluster, computed as the number of MCs per unit area (A).
5) *Scatteredness*: represented by the mean and the standard deviation of the interdistances between neighboring MCs; the neighbors are determined based on the Delaunay triangulation of the MCs as described above.
6) *Solidity*: computed as the ratio between cross sectional area A and the area of the convex hull formed by the MCs.
7) *Invariant moment* ϕ_1: a regional descriptor that is invariant to translation, rotation, or scaling [39].
8) *Moment signature*, as defined in [37]: a measure of boundary roughness, computed based on the distance deviation of a point on the boundary from the center of the region.
9) *Normalized Fourier descriptor*, also as defined in [37]: a frequency-domain characterization of the smoothness of the boundary.

These feature components (a total of 10, with 2 for scatteredness) were first computed for each MCC in the mammograms. All these feature components were then normalized to have the same dynamic range (0,1). Each MCC was then labeled with a feature vector $\mathbf{u}$ formed by these components. These feature vectors were paired with the observer similarity data to form the training and testing samples.

In summary, we have the following data set:
$$S = \{(\mathbf{x}_i, y_i), i = 1, 2, \ldots, 465\} \tag{20}$$
where $\mathbf{x}_i$ denotes the computed feature vector for the i-th MCC pair, and y_i is the observer SC of the pair. This set was used for the subsequent training and testing of the proposed framework.

D. Machine Training and Performance Evaluation

1) Preparation of Data Sets: For the first stage, the MCC pairs in set S in (20) were first divided into two classes: class 1 representing sufficiently similar pairs, and class 2 representing dissimilar pairs. We chose a threshold $T_1 = 4$ so that samples in S were labeled as class 1 if their SCs were larger than T_1; otherwise they were labeled as class 2. In short, we denote this set as
$$S_1 = \{(\mathbf{x}_i, d_i), i = 1, 2, \ldots, 465\} \tag{21}$$
where $d_i = \mathrm{sgn}(y_i - T_1)$. There were in total 229 samples in class 1, and 236 samples in class 2. This set was used subsequently to train and test the first-stage classifier.

For the regression stage, we chose only those pairs in S with SCs larger than T_1, i.e., those belonging to class 1 in set S_1. We denote this set as
$$S_2 = \{(\mathbf{x}_i, y_i) : y_i > T_1, i = 1, 2, \ldots, 465\}. \tag{22}$$

2) Performance Evaluation: For training and testing of the learning machines (both the classification stage and the regression stage), we applied the following cross-validation procedure [40]: 1) the images were selected in turn so that during each run only one image was chosen (as a query), based on which the data samples (S_1 or S_2) were divided into the following two sets: one for training, which consisted of all the samples not involving the chosen image, and the other for testing, which consisted of only those samples involving the chosen image; 2) in each run the learning machine (either classification or regression) was then trained using the resulting training set, and tested for performance using the testing set; 3) the test results were then averaged over all the different runs to obtain the generalization performance (e.g., classification error, retrieval precision, etc.).

To evaluate the performance of the retrieval network, we used the so-called *precision-recall curves* [1]. The retrieval *precision* is defined as the proportion of the images among all the retrieved that are truly relevant to a given query; the term *recall* is measured by the proportion of the images that are actually retrieved among all the relevant images to a query. The precision-recall curve is a plot of the retrieval precision vs the recall over a continuum of the operating threshold T (Fig. 1).

As the ground truth in calculation of the precision-recall curves, we considered an image to be truly relevant to a query if its corresponding observer SC is larger than a preselected threshold T_2. In our experiments, $T_2 = 6$ was used.

3) Relevance Feedback: To demonstrate the effect of relevance feedback, we performed the following experiments: for each query, the trained retrieval network was first applied to retrieve images from the database; among the images retrieved, the one with the highest SC (based on the pre-existing observer data) was chosen as the relevant feedback image (in case there was a tie, random selection was used to break the tie). The proposed relevance feedback procedure was then applied to retrieve a new set of images. The precision-recall curves were then computed based on this new set of images.

4) Impact of Parameters: To demonstrate the impact of various parameters involved in the training and testing of the proposed network on the overall performance, we also evaluated

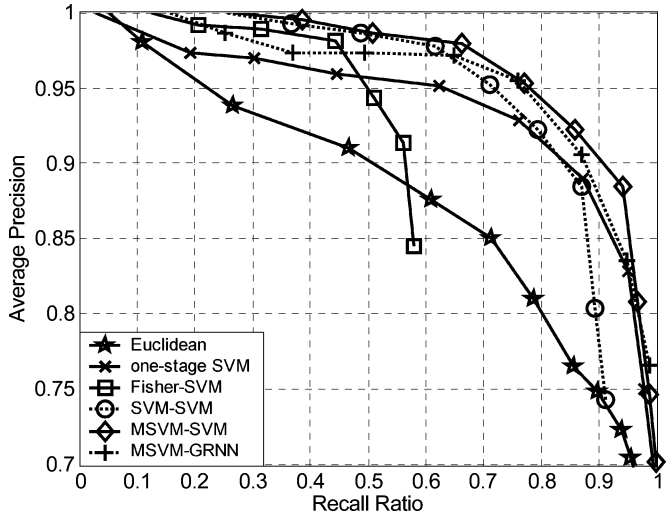

Fig. 6. Plot of precision-recall curves obtained from various network structures.

the precision-recall curves when these parameters were varied, including the thresholds T_1 and T_2, and the internal parameters of the trained learning machine.

VI. Experimental Results

The proposed two-stage learning approach was thoroughly tested and evaluated for retrieval under various learning-machine settings. We summarize the results in Fig. 6 using the precision-recall curves for the following different network structures:

1) a linear Fisher discriminant for the first stage and an SVM for the second stage (Fisher-SVM);
2) a linear SVM for the first stage and an SVM for the second stage (SVM-SVM);
3) a linear SVM with the modified objective function in (6) for the first stage and an SVM for the second stage (MSVM-SVM);
4) a linear SVM with the modified objective function in (6) for the first stage and a GRNN for the second stage (MSVM-GRNN).

In the first three structures a Gaussian kernel was used in the SVM for the second stage. We note that similar performance was also achieved when a polynomial kernel was used, of which the results are omitted for clarity of the plots.

For comparison, we also show in Fig. 6 the precision-recall curve obtained when a single stage SVM regression network was used for retrieval (SVM). In this case, the SVM with a Gaussian kernel was trained and tested directly using the samples formed from the entire set of observer SCs.

Moreover, we show in Fig. 6 the precision-recall curve obtained when a naïve Euclidian metric was used as the similarity measure. In this case, the images with features vectors closest to a query were retrieved.

From these results we see that the two-stage network (MSVM-SVM) achieves the best performance; and all the learning-based networks outperform that based on the Euclidian distance. Note that the precision-recall curves corresponding to both Fisher-SVM and SVM-SVM drop below that of the single-stage network (SVM) as the recall ratio is increased toward unity. This can be explained as follows: at a fixed operating threshold the first stage Fisher or SVM classifier discards some of the relevant images for a query with a nonzero probability, preventing the recall ratio from reaching 1 (as the retrieval threshold T gets decreased). The use of a modified SVM classifier in the first stage avoids this pitfall.

TABLE I
Parametric Settings of the Trained Retrieval Networks

Networks	Classification Stage	Regression Stage
Fisher-SVM	-	C=100, σ=1.5, ε =0.5
SVM-SVM	C=1000	C=100, σ=1.5, ε =0.5
MSVM-SVM	$C^+ = 10^5$, $C^- = 5\times10^4$	C=100, σ=1.5, ε =0.5
MSVM-GRNN	$C^+ = 10^5$, $C^- = 5\times10^4$	σ=0.25
SVM	-	C=100, σ=1.5, ε =0.5

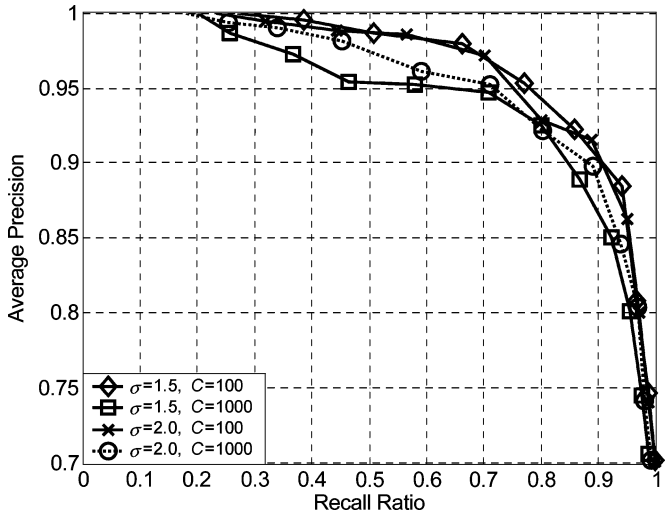

Fig. 7. Precision-recall curves obtained by the MSVM-SVM network when the parameters in the regression stage are varied from their tuned values of $\sigma = 1.5, C = 100$ in Table I to: (i) $\sigma = 1.5, C = 1000$, (ii) $\sigma = 2.0, C = 100$, and (iii) $\sigma = 2.0, C = 1000$.

Finally, the parametric settings for the learning machines corresponding to each of the network structures above are listed in Table I. In our experiments each network structure was studied over a wide range of parametric settings; the precision-recall curves in Fig. 6 represent the best results for each case. The performance of the learning networks was found to be considerably robust and insensitive to changes in the parameters (such as regularization parameter C, insensitive bound ε, kernel width σ, etc.). To demonstrate this, we show in Fig. 7 the resulting precision-recall curves for the case of MSVM-SVM when the parameters C and σ of the regression stage are varied from their tuned values listed in Table I.

In Figs. 8 and 9, we show some retrieval examples for two given query images. These results demonstrate that the two-stage network can indeed improve retrieval performance.

In Fig. 10, we show the precision-recall curves obtained using the weighted SC method for the two-stage MSVM-GRNN. Similar results were also obtained for SVM and MSVM-SVM (but not shown for brevity). As can be seen, the proposed feedback procedures can further improve the retrieval performance.

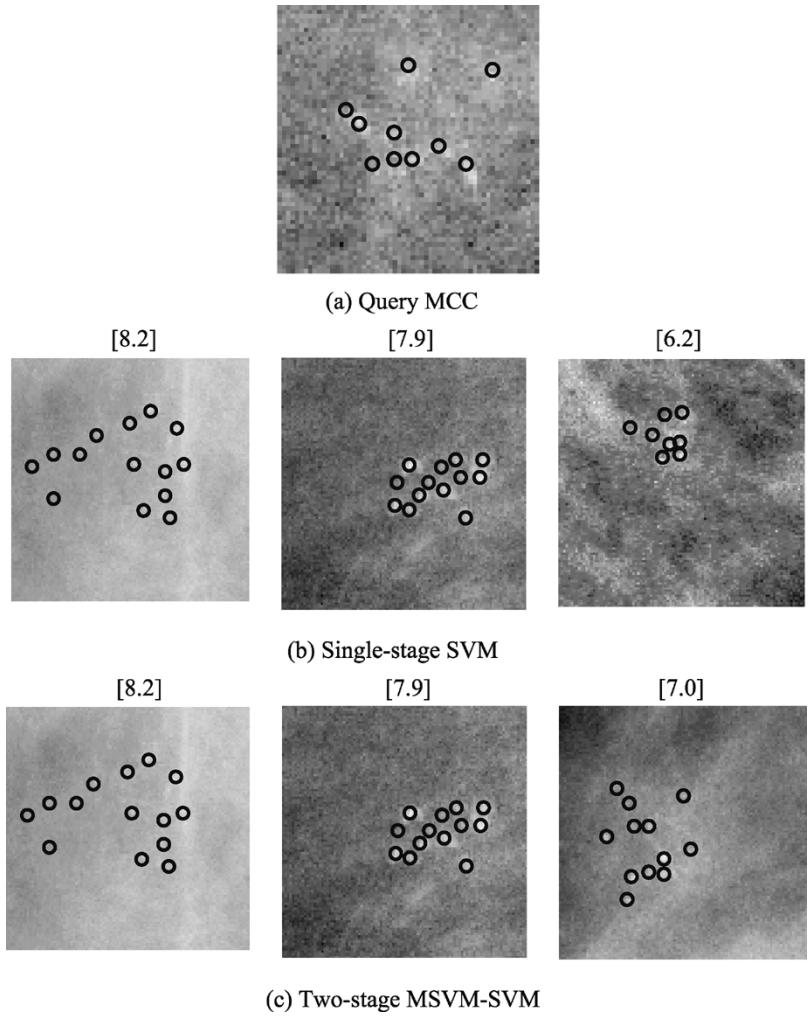

Fig. 8. Top three images retrieved by the single-stage SVM (b) and the two-stage MSVM-SVM (c), respectively, for a given query MCC (a). Numbers in brackets on top of each cluster are the user SCs.

In addition, we show in Fig. 11 the precision-recall curves obtained for MSVM-SVM when different values for the thresholds T_1 and T_2 were used. We note that varying the threshold T_1 leads to variations in the training datasets S_1 in (21) and S_2 in (22) for the two-stage network, and varying the threshold T_2 leads to variation in the ground truth for producing the precision-recall curves. These results demonstrate that the retrieval performance by the network is somewhat robust to these variations.

Finally, in Fig. 12 we show the average fraction of images among the top k retrieved images ($k = 1$, 2, 3, and 4) by MSVM-SVM for each query that actually match the disease condition of the query, obtained using the leave-one-out procedure. For comparison, we also show in Fig. 12 the matching fractions when the observer score (ground truth) is used. It can be seen that when the observer SCs were used, the average matching fraction was around 70%; the two-stage network could achieve a matching fraction above 60%. In particular, the most similar image (when $k = 1$) retrieved by the two-stage network can have a matching fraction as high as 76.7%, which is even higher than that of the observer SCs (66.7%). While this might seem surprising, such a results is possible because the regression network was trained based on similarity data from a wide range of SCs, and the resulting regression function has an inherent noise-smoothing effect (which can reduce the uncertainty in the observer data). We also conducted a binomial test [41] to establish the statistical significance of these results, as compared to what would have been achieved by random pairing (of which the expected matching fraction is 51.95%, determined by the distribution of the cases in the dataset); the p-value is 0.0034 for 76.7%, and 0.053 for 66.7%. Note that the disease information of the clusters had been kept unavailable during the observer study so that the observers were not influenced by the disease condition when scoring the similarity between MCCs.

In our experiments, the two-stage networks (MSVM-SVM and MSVM-GRNN) can provide 4 $\sim$ 5-fold reduction in computation time as compared to that of a single-stage network (SVM).

VII. Conclusion

In this paper, we have proposed a learning machine-based framework for modeling human perceptual similarity for content-based image retrieval. The proposed approach was developed and evaluated for retrieval of clinical mammograms containing clustered microcalcifications. The results demonstrated that a learning framework can be used effectively to

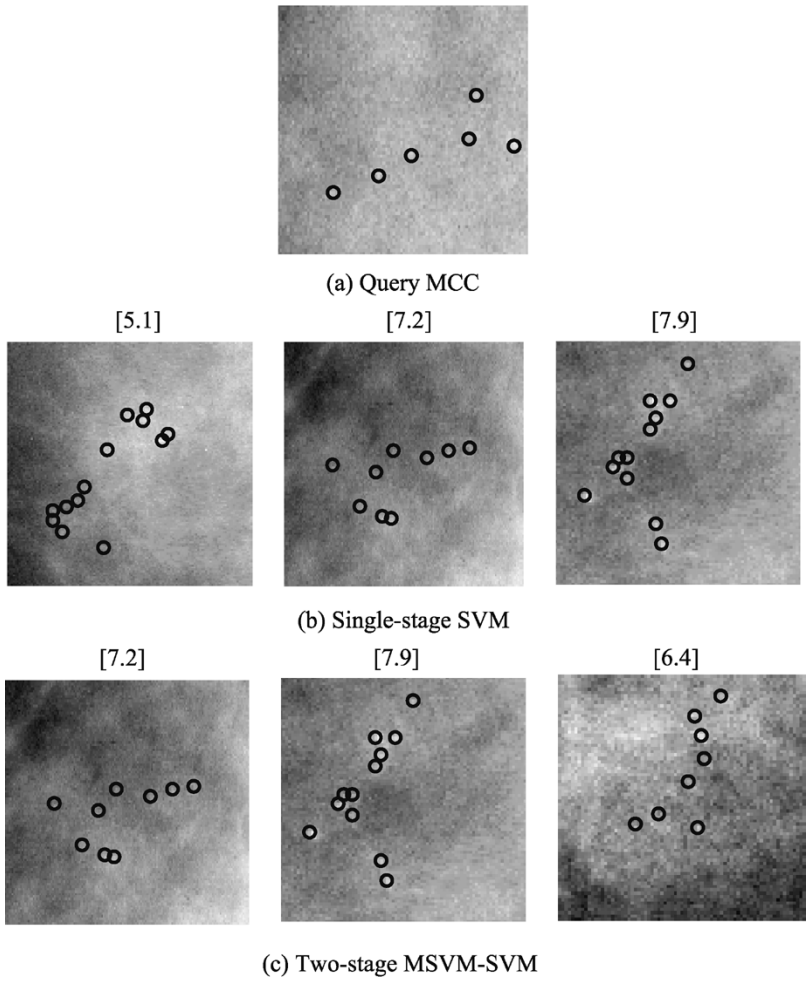

Fig. 9. Top three images retrieved by the single-stage SVM (b) and the two-stage MSVM-SVM (c), respectively, for a given query MCC (a). Numbers in brackets on top of each cluster are the user SCs.

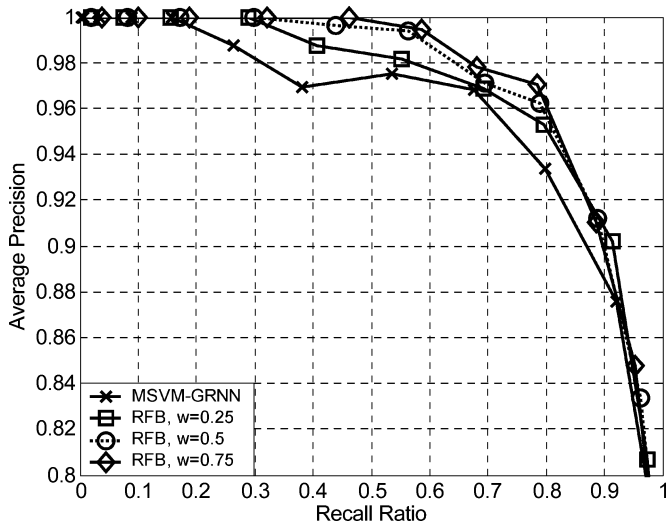

Fig. 10. Precision-recall curves using relevance feedback (RFB) for the two-stage network MSVM-GRNN.

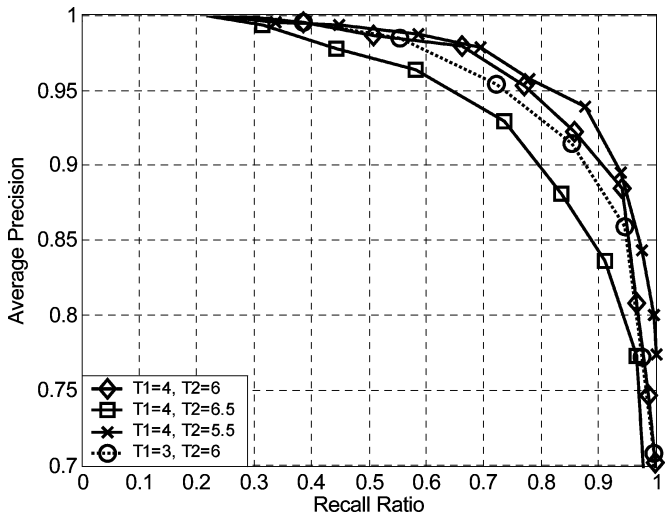

Fig. 11. Precision-recall curves obtained by the MSVM-SVM network when different values were used for the thresholds T_1 and T_2: i) $T_1 = 4, T_2 = 6.5$; ii) $T_1 = 4, T_2 = 5.5$; and iii) $T_1 = 3, T_2 = 6$.

model the perceptual similarity, thereby serving as basis for retrieving visually similar mammograms from a database. It was demonstrated that a hierarchical two-stage learning network can offer several advantages over a single-stage one, including faster speed and retrieval accuracy. Furthermore, the use of relevance feedback in such a framework can be used to further improve the retrieval performance. In our future work we will explore the use of incremental learning to adapt the learning network online to a user's feedback; we will also investigate

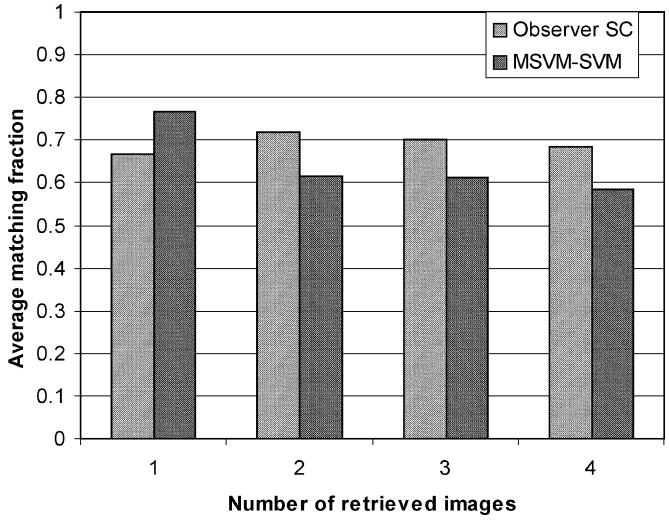

Fig. 12. Average fraction of images among the top k retrieved images ($k = 1$, 2, 3, and 4) that actually match the disease condition of the query. For comparison, the matching fraction is also shown when the observer score (ground truth) is used.

the clinical benefit of using the developed retrieval framework for computer-aided diagnosis.

ACKNOWLEDGMENT

R. M. Nishikawa is a shareholder in R2 Technology, Inc., Sunnyvale, CA.

REFERENCES

[1] A. D. Bimbo, *Visual Information Retrieval*. San Mateo, CA: Morgan Kauffman Publishers, 1999.
[2] Y. Rui and T. Huang, "Image retrieval: current techniques, promising directions and open issues," *J. Visual Commun. Image Representation*, vol. 10, pp. 39–62, 1999.
[3] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A sketch retrieval method for full color image database," in *Proc. 11th Int. Conf. Pattern Recogn.*, The Hague, The Netherlands, 1992, pp. 530–533.
[4] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, and C. Faloutsos, "The QBIC project: querying images by content using color, texture, and shape," *SPIE Proc. Storage Retrieval Image Video Databases*, vol. 1908, pp. 137–146, 1993.
[5] S. Wong, "CBIR in medicine: still a long way to go," in *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries*, Santa Barbara, CA, June 1998, p. 115.
[6] H. 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. M. Kelly and T. M. Cannon, "CANDID: comparison algorithm for navigating digital image databases," in *Proc. Int. Working Conf. Scientific and Statistical Database Management*, 1994, pp. 252–258.
[8] A. Guimond and G. Subsol, "Automatic MRI database exploration and applications," *Pattern Recogn. Artif. Intell.*, vol. 11, no. 8, pp. 1345–1365, 1997.
[9] Y. Liu and F. Dellaert, "A classification based similarity metric for 3D image retrieval," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 1998, pp. 800–805.
[10] J. Sklansky, E. Tao, M. Bazargan, C. Ornes, R. Murchison, and S. Teklehaimanot, "Computer-aided, case-based diagnosis of mammographic regions of interest containing microcalcifications," *Academic Radiol.*, vol. 7, no. 6, pp. 395–406, 2000.
[11] C. Ornes and J. Sklansky, "A visual neural classifier," *IEEE Trans. Systems, Man, and Cybernetics*, pt. B, vol. 28, pp. 620–625, 1998.
[12] Y. Kawata, N. Niki, H. Ohmatsu, M. Kusumoto, R. Kakinuma, K. Mori, H. Nishiyama, K. Eguchi, M. Kaneko, and N. Moriyama, "Three-dimensional CT image retrieval in a database of pulmonary nodules," in *Proc. of IEEE International Conference on Image Processing*, Rochester, NY, Sept. 2002, pp. 149–152.
[13] I. El-Naqa, M. N. Wernick, Y. Yang, and N. P. Galatsanos, "Image retrieval based on similarity learning," in *Proc. IEEE Int. Conf. Image Processing*, Vancouver, BC, Canada, 2000, pp. 722–725.
[14] A. I. Mushlin, R. W. Kouides, and D. E. Shapiro, "Estimating the accuracy of screening mammography: a meta-analysis," *Amer. J. Preventive Med.*, vol. 14, no. 2, pp. 143–153, 1998.
[15] R. N. Strickland and H. L. Hahn, "Wavelet transforms for detecting microcalcifications in mammograms," *IEEE Trans. Med. Imag.*, vol. 15, no. 2, pp. 218–229, 1996.
[16] E. A. Sickles, "Mammographic features of 300 consecutive nonpalpable breast cancers," *Amer. J. Roentgenol.*, vol. 146, pp. 661–663, 1986.
[17] D. B. Kopans, "The positive predictive value of mammography," *Amer. J. Roentgenol.*, vol. 158, pp. 521–526, 1992.
[18] J. G. Elmore, C. K. Wells, C. H. Lee, D. H. Howard, and A. R. Feinstein, "Variability in radiologists' interpretations of mammograms," *New Engl. J. Med.*, vol. 331, no. 22, pp. 1493–1499, 1994.
[19] R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computer aided detection of clustered microcalcifications in digital mammograms," *Med. Biological Eng. Computing*, vol. 33, pp. 174–178, 1995.
[20] P. L. Miller, "Critiquing anesthetic management: the "ATTENDING" computer system," *Anesthesiology*, vol. 58, pp. 362–369, 1983.
[21] B. Ripley, *Pattern Recognition Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[22] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
[23] M. Pontil and A. Verri, "Support vector machines for 3-D object recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 637–646, June 1998.
[24] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification," in *Proc. IEEE Workshop Neural Networks for Signal Processing*, Sydney, Australia, Dec. 2000, pp. 775–784.
[25] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: application to face detection," in *Proc. Computer Vision and Pattern Recognition*, Puerto Rico, 1997, pp. 130–136.
[26] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. Int. Conf. Machine Learning*, Slovenia, June 1999.
[27] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
[28] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, "A support vector machine for approach for detection of microcalcifications," *IEEE Trans. Med. Imag.*, vol. 21, pp. 1552–1563, Dec. 2002.
[29] D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Network*, vol. 2, pp. 568–576, Nov. 1991.
[30] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, Sept. 1998.
[31] I. El-Naqa, Y. Yang, N. P. Galatsanos, and M. N. Wernick, "Relevance feedback based on incremental learning for mammogram retrieval," in *Proc. IEEE Int. Conf. Image Processing*, Barcelona, Spain, Sept. 13–17, 2003.
[32] *Illustrated Breast Imaging Reporting and Data System*, 3rd ed. Reston, VA: American College of Radiology, 1998.
[33] M. G. Kendall, *Rank Correlation Methods*, 4th ed. London, U.K.: Griffin, 1970.
[34] J. H. Zar, *Biostatistical Analysis*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[35] Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: automated feature analysis and classification," *Radiology*, vol. 198, pp. 671–678, 1996.
[36] Y. Jiang, R. M. Nishikawa, R. Schmidt, C. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer aided diagnosis," *Acad Radiol*, vol. 6, pp. 22–33, 1999.
[37] S. Liang, R. Rangayyan, and J. E. Desautels, "Application of shape analysis to mammographic calcifications," *IEEE Trans. Med. Imag.*, vol. 13, pp. 263–274, June 1994.
[38] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 2nd ed. New York: Academic, 2003.
[39] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Reading, MA: Addison-Wesley, 1992.
[40] R. Kennedy, Y. Lee, B. Van Roy, C. D. Reed, and R. P. Lippman, *Solving Data Mining Problems Through Pattern Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1998.
[41] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*. New York: Wiley, 1972.