Unsupervised Feature Selection Applied to Content-Based Retrieval of Lung Images

Jennifer G. Dy, *Member*, *IEEE*,
Carla E. Brodley, Avi Kak, *Member*, *IEEE*,
Lynn S. Broderick, and
Alex M. Aisen

Abstract—This paper describes a new hierarchical approach to content-based image retrieval called the "customized-queries" approach (CQA). Contrary to the single feature vector approach which tries to classify the query and retrieve similar images in one step, CQA uses multiple feature sets and a two-step approach to retrieval. The first step classifies the query according to the class labels of the images using the features that best discriminate the classes. The second step then retrieves the most similar images within the predicted class using the features customized to distinguish "subclasses" within that class. Needing to find the customized feature subset for each class led us to investigate feature selection for unsupervised learning. As a result, we developed a new algorithm called FSSEM (feature subset selection using expectation-maximization clustering). We applied our approach to a database of high resolution computed tomography lung images and show that CQA radically improves the retrieval precision over the single feature vector approach. To determine whether our CBIR system is helpful to physicians, we conducted an evaluation trial with eight radiologists. The results show that our system using CQA retrieval doubled the doctors' diagnostic accuracy.

Index Terms—Image retrieval, feature selection, clustering, expectation-maximization, unsupervised learning.

1 Introduction

CONTENT-BASED image retrieval (CBIR) refers to the ability to retrieve images on the basis of image content, as opposed to on the basis of some textual description. In radiology, diagnosis of high resolution computed tomography (HRCT) images of the lungs is particularly well placed to benefit from CBIR. There are two reasons. First, there is often a substantial difference in the ability of specialists and nonspecialists to diagnose lung disease. Second, for radiologists who do not frequently interpret HRCT scans, common practice is to use a reference text such as [35] to find images that are similar to the query image. Because the images in the text have known diagnoses, the radiologist can base diagnosis on their similarity to the images of the patient at hand. In the medical domain, the goal of a CBIR system is to aid doctors to diagnose a patient by retrieving images with known pathologies that are similar to the patient's image(s).

The traditional approach to CBIR represents each image in the database by a vector of feature values [12], [20], [23], [28], [31], [32]. During retrieval, the images that are most similar to the query in terms of some distance measure (e.g., Euclidean distance) are then retrieved. We call this approach to CBIR the *single feature vector approach*. For this approach, the choice of similarity metric and features to include for characterizing the images are critical factors in its ability to achieve high retrieval precision.

This raises the question of how to define similarity. Two images are similar if they are judged similar by the human user. Defining a similarity metric is difficult because human perception is difficult to model. Furthermore, similarity varies from user to user and from one context to another. However, in medical domains and HRCT of the lungs in particular, similarity means images that correspond to the same disease type, stage, severity, and treatment. We observed that the single feature vector approach will not perform well for this domain because *the features that are most effective in discriminating among images from different classes may not be the most effective for retrieval of images belonging to the same subclass within a class*. This occurs for domains in which not all pairs of images within one class have equivalent perceptual similarity (i.e., a hierarchy of classes exists).

In HRCT of the lungs, a query image may differ from other images within the same disease class on account of the severity of disease and other such factors. Fig. 1 illustrates this point. Notice that, within the class Centrilobular Emphysema (CE), Fig. 1d is visually dissimilar to Figs. 1b and 1c. This dissimilarity corresponds to differences in disease severity. Furthermore, the features that best discriminate among disease classes are different from the features that best discriminate between subclasses within a disease class. For example, a feature that distinguishes Paraseptal Emphysema from CE is the distance of the "pathology bearing region (PBR)"[1] from the boundary of the lung, whereas the features that best discriminate the images within class CE are those that measure the gray level intensity of the PBR.

To handle this problem, we have designed, implemented, and evaluated an approach called the "customized-queries" approach (CQA) that uses a two-level approach to retrieval. First, it classifies the query according to disease class labels using the features that best discriminate the classes. Then, it retrieves the most similar images using the features customized to distinguish "subclasses" within a disease class. Because we are not provided with labels indicating the subclasses within a disease, we must learn them by applying unsupervised learning methods. Ideally, our clustering algorithm would find clusters within a disease class corresponding to disease stage, severity, and/or treatment. In practice, we will seek clusters that define homogenous groupings with respect to the features chosen to calculate visual similarity.

Forming a hierarchy of features for retrieval and storage has been explored by other researchers, but their end goals for doing so differ from ours. For example, in the FourEyes system [19], highly structured objects in images, such as buildings and trees, are represented hierarchically to facilitate structural comparisons with a query image. Ma and Manjunath [17] built a hybrid neural network classifier to classify the query as one of the given classes and, then, select the n most similar images within that class cluster using Euclidean distance. Note that the same feature set is used both for classification and for retrieval after classification. Chen and Bouman [3] developed an approach that organizes images in "similarity pyramids" by grouping together images that are closest in L1 distance. The resulting organization is used for indexing and browsing purposes. CQA differs from these approaches because it uses different feature sets for comparing similarity at each level and for each class. Our approach is not limited in applicability to medical domains, but can be applied to any domain where the features that best discriminate the given classes are different from those that characterize subclasses within a class.

1. The PBR is the region marked by the physician as the region of interest or diseased region. The PBR's are encircled by a white boundary as shown in Fig. 1.

- *J.G Dy is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115. E-mail: jdy@ece.neu.edu.*
- *C.E. Brodley and A. Kak are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907. E-mail: {brodley, kak}@ecn.purdue.edu.*
- *L.S. Broderick is with the Department of Radiology, University of Wisconsin, Madison, WI 53706. E-mail: lsbroderick@facstaff.wisc.edu.*
- *A.M. Aisen is with the Department of Radiology, Indiana University Medical Center, Indianapolis, IN 46202. E-mail: aaisen@iupui.edu.*

Manuscript received 12 June 2001; revised 26 Feb. 2002; accepted 21 July 2002.
Recommended for acceptance by C. Schmid.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 114338.

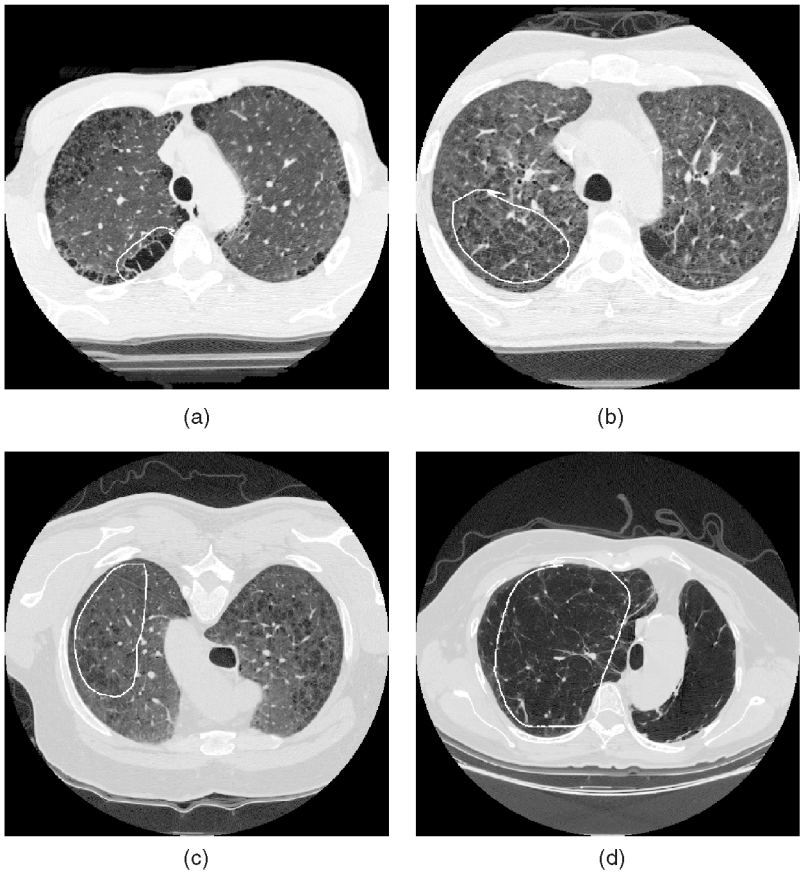

Fig. 1. (a) Paraseptal Emphysema image. (b) Centrilobular Emphysema (CE) image in subclass 1. (c) CE image in subclass 1. (d) CE image in subclass 2.

2 The Customized-Queries Approach (CQA)

We refer to our approach as the *customized-queries* approach because it allows the system to "customize" the feature set to the query image's class. By customization, we mean that only those features that retrieve the most similar images within a disease class are used. CQA, applied to our HRCT-lung domain, has two levels of similarity. The first level addresses the similarity of the query image to each disease class. The second level addresses similarity in terms of disease severity, stage, treatment, structure, and/or visual appearance. Section 2.1 describes how we customize the Level 1 features, and Section 2.2 describes customization of Level 2 features.

During retrieval, a radiologist selects a query image and then marks the suspected diseased region (PBR). CQA, then, classifies the query into one of the disease categories using our Level 1 features and classifier. Next, CQA retrieves n (default four) similar images within the query's class, utilizing the Level 2 feature subset customized to the query's classified disease class. We apply Euclidean distance as our dissimilarity metric for retrieval with each customized feature standardized to variance one.

2.1 Level 1: Feature Selection for Classification

The first step in CQA is to classify the query into one of the given disease classes. These pathology class labels are confirmed diagnoses obtained from medical records, hence, we can consider them as ground truth labels. We would like to find the features that most accurately discriminate among the disease classes and will yield the highest classification accuracy. To this end, we compared several classifiers and found that the most accurate was formed by C5.0, an algorithm for forming boosted rule sets from decision trees [21]. Boosting is an algorithm that improves classification accuracy by utilizing multiple classifiers [13], [25].

2.2 Level 2: Feature Selection for Unsupervised Learning

For the first level of CQA, disease classes are available to guide the search for the features that best discriminate among the disease classes. For the second level, we wish to find the features that best define similarity within a single disease class. Because we are not provided with subclass labels within each disease class, we are faced with the problem of feature selection for unsupervised (unlabeled) data.

Although research in feature selection for supervised learning has a long history [1], [2], [14], [16], research in feature selection for unsupervised learning (clustering) is relatively new. Devaney and Ram [5] and Talavera [30] developed feature selection algorithms for COBWEB [11] (a hierarchical clustering algorithm). Vaithyanathan and Dom [34] formulated an objective function for choosing the feature subset and finding the optimal number of clusters for a document clustering problem using a Bayesian statistical estimation framework with each cluster modeled as a multinomial. These methods are not suitable for our application because we prefer a partitional clustering algorithm instead of hierarchical, and the multinomial model used by Vaithyanathan and Dom [34] is not ideal for our type of real-valued features.

Here, our goal in feature selection is to find a minimum set of features that best discriminates the subclasses. Since the subclasses are unknown, we need to find the features and uncover the subclasses simultaneously. We introduce our method, FSSEM (feature subset selection using EM clustering), which is inspired by the wrapper approach to feature subset selection for supervised learning [16]. Instead of using feature subset selection wrapped around a classifier, we wrap it around a clustering algorithm. The basic idea of our approach is to search through feature subset space, evaluating each subset F_t, by first clustering in space F_t using EM clustering and, then, evaluating the resulting clusters in space F_t using our chosen feature selection criterion. The result of

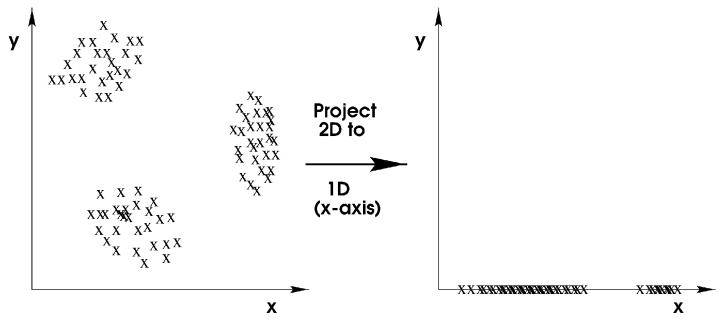

Fig. 2. The number of cluster components varies with dimension.

this search is the feature subset that optimizes our chosen criterion function. In [7], we illustrated that FSSEM is an effective feature selection for unsupervised learning algorithm and evaluated various choices for how to implement unsupervised feature selection within the wrapper framework. In the remainder of this section, we introduce our design choices for the domain of clustering HRCT images of the lungs.

Search Method. Because there are 2^d feature subsets, where d is the number of available features, exhaustive search is intractable. Our current implementation applies sequential forward selection (SFS) [14] to search the features. One could choose other search methods in the wrapper framework such as sequential backward elimination, forward-backward, sequential floating searches, or genetic algorithms [15], [24]. SFS is a greedy search algorithm that adds one feature at a time. This method adds the feature that, when combined with the current chosen set, yields the largest improvement to our feature selection criterion. SFS does not guarantee an optimal solution, but it is simple with $O(d^2)$ complexity and is sufficient for our purpose.

Clustering Algorithm. Clustering aims to find the natural groupings of the data. In our application, we assume that each subclass grouping is Gaussian and use EM clustering. We define EM clustering as the expectation-maximization (EM) [4] of a finite multivariate Gaussian mixture [33], [36]. We apply the EM algorithm to estimate the maximum likelihood mixture model parameters and the cluster probabilities of each data point. EM clustering results in "soft" clusters (i.e., each data point belongs to every cluster with some probability). See [18] for a complete description of the EM algorithm applied to clustering.

In the EM algorithm, we start with an initial estimate of our parameters and, then, iterate using the update equations until convergence. *Note that EM is initialized for each new feature subset.* The EM algorithm can get stuck at a local maximum far from the actual solution, hence, the initialization values are important. To initialize EM, we used the subsampling initialization algorithm proposed by Fayyad et al. [10] (with 10 percent subsampling and $J = 10$ subsampling iterations). We then iterate until convergence (likelihood does not change by 0.0001) or up to n (default 500) iterations; whichever comes first. EM estimation is constrained away from singular solutions in the parameter space by limiting the diagonal elements of the component covariance matrices Σ_j to be greater than $\delta = 0.000001\sigma^2$, where σ^2 is the average of the variances of the unclustered data. Adding the identity matrix multiplied by a small scalar (αI) to a matrix where $\alpha > 0$ makes the final matrix positive definite (i.e., all eigenvalues are greater than zero and hence nonsingular).

Feature Evaluation Criterion. We evaluate the clusters discovered by our candidate feature subset with our feature evaluation criterion. We would like this criterion to measure how well the candidate feature subset separates the clusters (i.e., subclasses). One measure is the $trace(S_w^{-1}S_b)$ criterion used in discriminant analysis [14]. S_w measures how scattered the samples are from their cluster means (compactness). S_b measures how scattered the cluster means are from the total mean (separability). We would like the distance between each pair of samples in a particular cluster to be as small as possible and the cluster means to be as far apart as possible with respect to the chosen similarity metric. $S_w^{-1}S_b$ is S_b normalized by the average cluster covariance. Hence, the larger the value of $trace(S_w^{-1}S_b)$ is, the larger the normalized distance between clusters is, which results in better cluster discrimination. S_w and S_b are defined as follows:

$$S_w = \sum_{i=1}^{k} \pi_i E\{(X-\mu_i)(X-\mu_i)^T/\omega_i\} = \sum_{i=1}^{k} \pi_i \Sigma_i$$
$$S_b = \sum_{i=1}^{k} \pi_i (\mu_i - M_o)(\mu_i - M_o)^T$$
$$M_o = E\{X\} = \sum_{i=1}^{k} \pi_i \mu_i,$$

where π_i is the probability of class ω_i, X is a random feature vector representing the image, μ_i is the mean vector of class ω_i, M_o is the total mean across all data points or images in the database, ω_i is the class ω_i, Σ_i is the covariance matrix of class ω_i, and $E\{\cdot\}$ is the expected value operator. Among the many possible separability criteria, we choose $trace(S_w^{-1}S_b)$ as our criterion because it is invariant under any nonsingular linear transformation [14].

Issue 1: Finding the Number of Clusters. We tackle two issues that arise in applying our wrapper approach to feature selection for unsupervised learning: the need to 1) find the number of clusters and 2) normalize the feature selection criteria with respect to dimension. In [6] and [7], these issues are discussed in detail.

The number of clusters, k, *depends on the feature subset*. Fig. 2 illustrates this point. In two dimensions (shown on the left), there are three clusters, whereas in one-dimension (shown on the right), there are only two clusters. It is not a good idea to use a fixed k in feature search, because different feature subsets require different numbers of clusters. And, using a fixed number of clusters for all subspaces does not model the data in each respective subspace correctly. Thus, we need to find the number of clusters while clustering each candidate feature subset. To search for k for a given feature subset, we add a minimum description length [22] penalty term to the log-likelihood criterion. Our new objective function becomes: $F(k,\Phi) = \log(f(X|\Phi)) - \frac{1}{2}L\log(Nd)$, where N is the number of data points, d is the dimension, L is the number of real numbers needed to specify the parameters Φ, and $\log(f(X|\Phi))$ is the log-likelihood of our observed data X given the parameters Φ. Note that L and Φ vary with k. A penalty term is needed because the maximum likelihood estimate increases as more clusters are used. Without the penalty, the likelihood is at a maximum when each data point is considered as an individual cluster. There are myriad ways to find the "optimal" number of clusters k with EM clustering (see Smyth [29] for an overview).

Issue 2: Criterion Normalization. In order to compare feature subsets of different dimensionality, we need to normalize our feature selection criterion. The separability criterion is biased toward higher dimensions. The criterion value monotonically increases as features are added assuming equal clustering assignments [14]. This is not desirable because we would like to keep the minimum number of features needed. To normalize our criterion function, we project the clusters found to the two feature subspaces

we are comparing. For example, we would like to see whether subset S_2 leads to better clusters than subset S_1. We refer to the clusters found in S_2 as C_2, and those found in S_1 as C_1. We normalize the criterion value for S_1 as:

$$normalizedValue(S_1) = CRIT(S_1, C_1) \cdot CRIT(S_2, C_1).$$

And, the criterion value for S_2 as:

$$normalizedValue(S_2) = CRIT(S_2, C_2) \cdot CRIT(S_1, C_2).$$

If $normalizedValue(S_1) > normalizedValue(S_2)$, we choose feature subset S_1. When the normalized criterion values are equal for S_1 and S_2, we favor the lower dimensional feature subset. $CRIT(S_2, C_2)$ is the ordinary criterion value. And, $CRIT(S_1, C_2)$ is the criterion value of C_2 projected to S_1. $CRIT(S_1, C_1)$ is the criterion value of C_1 in its original feature space. $CRIT(S_2, C_1)$ is the criterion value of the projection of C_1 on S_2. Since we project our clustering assignments to the two feature subsets we are comparing, after normalization, we are now comparing criteria in the same dimensions.

Complexity. The overall complexity of FSSEM is $O(d^4k^2ne)$, where d is the number of features, k is the maximum number of clusters, n is the number of data points, and e is the average number of EM iterations. Note that finding Level 1 and Level 2 features is performed offline. These preselected feature subsets are available to the system during online retrieval.

3 Experiments

We present the results of two experiments. In the first experiment, we investigated whether CQA achieves better retrieval precision than the single feature vector approach. In the second experiment, we performed an evaluation trial with eight doctors to determine whether our system helped improve their diagnostic accuracy.

3.1 Experiment 1: CQA versus the Single Feature Vector Approach

In this experiment (conducted in Fall of 1999), the database consists of 312 HRCT lung images from 62 patients [8]. These images yield 615 pathology bearing regions (PBR's) [26], which are local image regions marked by the physician as pathological. A single image may have several PBR's and these PBR's may have different diagnoses. Throughout the experiment, we considered each PBR as a data point (i.e., a single image with three PBR's gives us three data points). We used 125 implemented features as our set of candidate features. A complete description of the full set of features is given in [9], [26]. The features include measures of geometric properties (centroid, area, distance from boundary), gray-level mean, standard deviation, gray-level histogram, area histogram, texture, and edginess measures of the local pathology bearing regions and of the global lung image [26]. Compared to the list of features in [9], [26], some features were removed because they were redundant. During our searches for the Level 2 features, we excluded the PBR location and size features that may capture systematic effects of the PBR markings made by our physician. Although these features cluster the images well, the resulting clustering does not group the PBR's according to visual similarity. This cuts down our initial feature space to 110 features. Our experiments compare the following two methods:

The single feature vector approach. The customized Level 1 features are used for retrieval across the entire database using Euclidean distance.

The customized-queries approach. The customized Level 1 features classify the query image as one of the Level 1 classes. CQA then retrieves the nearest neighbors within that class as measured by Euclidean distance over the customized Level 2 features corresponding to the predicted class.

In assessing the performance of CQA, we assumed an ideal classifier was used to classify the query as a Level 1 class. We did this to isolate the effect of the Level 1 classifier on using the appropriate Level 2 features in retrieving the images. This way, whenever a different Level 1 classifier is applied, the approximate effective retrieval precision can be computed as: Level 1 classifier accuracy times Level 2 retrieval precision.

In this experiment, we customize our Level 1 features by applying a wrapper approach that uses SFS to search the feature space wrapped around a one nearest-neighbor (1-NN) classifier [14]. We use a nearest-neighbor classifier because this is the classifier/retriever of the single feature vector approach. That way, we optimize the features for this approach. We chose 1-NN because, in a comparison to 2, 3, 4, and 5-NN, 1-NN yielded the smallest classification error. To estimate the classification error, we apply 10-fold cross-validation, which randomly partitions the data set into 10 mutually exclusive subsets. Classification error is computed with each partition (or fold) as the test set and the rest as the training set. The wrapper approach chose 11 Level 1 features, which is a substantial reduction from using all of 125 possible features. Table 1 presents a list of the selected Level 1 features, Level 2 features, and the number of clusters found for each disease class.

To determine which method is better, a radiologist in our team (Dr. Lynn Broderick) who specializes in lung pathology, was asked to evaluate the retrieval results of the two methods. Throughout the test, the radiologist was not informed as to which method produced the retrieved images. We used randomly selected images with the following class distribution: 18 from the C-Emphysema class, three from P-Emphysema, two from IPF, and one from each of EG, Bronchiectasis, Sarcoid, and Aspergillus as test query images. This test set is designed to have similar distribution proportions as our training data. Our database disease class distribution is available in [9]. The four images ranked most similar to the query image were retrieved for each method. Note that all images of the query patient are excluded from the search. The radiologist can choose from five responses: strongly-agree (SA), agree (A), not sure (NS), disagree (D), and strongly-disagree (SD) for each retrieved image. To measure the performance of each method, the following scoring system was used: 2 for strongly agree, 1 for agree, 0 for not sure, -1 for disagree, and -2 for strongly disagree. The results are summarized in Table 2.

The single vector approach received a total of -37 points, and CQA garnered 178 points. To evaluate the performance of the different methods, we report their retrieval precision in addition to our scoring system. Retrieval precision is the number of relevant retrievals divided by the total number of images retrieved. We cannot measure recall because we do not have the subclass labels. We considered SA and A as positive retrievals and NS, D, and SD as negative retrievals. The single vector approach resulted in 38.89 percent retrieval precision. There were many cases where the radiologist did not mark SA or A even though the retrieved images had the same diagnosis as the query image. Although those images belong to the same disease class, they have different disease severity, structure, and/or visual appearance. CQA resulted in 90.74 percent precision for an ideal Level 1 classifier. The Level 1 test accuracy for a 1-NN classifier is 80.65 percent. This leads to an approximate effective retrieval precision of 73.18 percent for CQA. From these results, we can see that customized queries dramatically improve retrieval precision compared to the single feature vector approach.

Our results show that it is not sufficient to retrieve images based on just the disease class. In addition, we need to find the best image within that class on the basis of visual similarity. Moreover, as shown in Table 1, the feature set that best discriminates the disease classes is different from the feature set that best discriminates between similar subclasses within each disease class. Hence, there exists a need for customized queries.

3.2 Experiment 2: Diagnosis with and without CBIR

We performed an evaluation trial in the Fall of 2000 to determine whether our CBIR system helps doctors (experts and nonexperts)

TABLE 1
The Level 1 Features, Level 2 Features, and the Number of Clusters Found

Level 1 Features	global gray level mean, global area, 2 global histogram features, closest fissure, distance from fissure, 2 texture constrast features, edge feature, local histogram feature, texture homogeneity	
Level 2 Features		
Disease Class	Feature Set	Clusters
Aspergillus	2 local gray histogram features	5
Bronchiectasis	global histogram, local gray mean, 2 local area histogram, 2 area histogram, 3 local histogram, 2 texture energy, texture correlation at distance 3	2
C-Emphysema	area histogram, local histogram, 2 texture energy, texture correlation	5
EG	global histogram, local histogram	4
IPF	global histogram, area histogram, 3 texture homogeneity	3
P-Emphysema	global histogram, texture homogeneity	4
Sarcoid	mean - global mean	5

diagnose new patients. The trial employed a custom coded Web-based interface which participants can access from any Internet-attached computer using a standard Java-enabled browser. The trial consisted of two phases:

Phase A. For each patient, we presented the doctor with one lung image (chosen to be representative of the disease by our lung specialist, Dr. Lynn Broderick) and with a list of diagnoses (including the choices "normal" and "other"). The physician was required to choose one and only one diagnosis, which the system records. It should be noted that the nature of the presentation was such that even an expert cannot achieve a "perfect" score. Only one image was presented for each patient, and the image was displayed showing a lung window only; the window and level could not be adjusted. No demographic or clinical data was provided. Furthermore, there was overlap in the list of possible diagnoses, that even an expert could not be expected to resolve without additional information.

Phase B. One week later we ran the same trial, but this time we allowed the doctors to use our system to assist with diagnosis. In this phase, the images displayed in phase A were again presented, one case at a time, and in a random order. The doctor now used the computer mouse to mark a suspected pathology bearing region in the query image. The system then retrieved the four most similar images with their corresponding disease labels. The system used the customized queries method for retrieval. The doctor was informed about the operation of the system and understood that the matching images chosen by the computer may or may not be correct; it was up to the physician to determine whether or not to accept the diagnosis presented by the computer. After viewing the displayed matching images, the physician chose the single "best" diagnosis from the same pull down list as in phase A.

Eight doctors participated in this trial: two lung experts, one liver expert, three general practitioners, one resident, and one fellow. We divided the data into train and test sets. Our retrieval method, CQA, was trained offline with 874 images from 149 patients resulting in 1,545 pathology bearing regions. A separate test set from 23 patients and 29 images was used as the query database. We designed our test set to have similar disease class proportion as the training data. The actual patients/images were selected randomly. In this experiment, we have a total of 210 base features (the original 125 plus Shyu's perceptual features [27]). In this experiment, we used C5.0, described in Section 2.1, as our Level 1 classifier.

The results of the trial show that our CBIR system improved the diagnosis accuracy for all types of physicians in our experiment, and helped nonspecialists more than specialists. Overall, our system improved diagnosis accuracy from 30.2 percent in phase A to 63.4 percent in phase B. The number of diagnoses that changed from wrong to right is 80, from right to wrong is three, and from wrong to wrong is 46. The low diagnosis accuracy in phase A can be attributed to the following four factors:

1. Doctors are used to looking at HRCT images through film.
2. Typically they have several image slices to view.

TABLE 2
Experiment 1: Results

Disease	Single Vector					CQA				
Class	SA	A	NS	D	SD	SA	A	NS	D	SD
CE	28	9	5	2	28	69	2	1	0	0
PE	0	0	4	0	8	10	0	1	0	1
IPF	5	0	0	0	3	3	2	2	0	1
EG	0	0	0	0	4	4	0	0	0	0
Sar.	0	0	0	0	4	0	0	0	0	4
Asper.	0	0	0	0	4	3	1	0	0	0
Bron.	0	0	0	0	4	3	1	0	0	0
total	33	9	9	2	55	92	6	4	0	6
score	-37					178				
ret. precision	38.89%					90.74%				

3. The patient's medical history is available.
4. Many of the doctors that participated are not lung specialists.

Nevertheless, provided with only a single HRCT image slice displayed on the computer, our system enabled the physicians to double their diagnostic accuracy.

4 Summary

We introduced the customized-queries approach to CBIR, which first classifies a query using the features that best differentiate the Level 1 classes and, then, customizes the query to that class by using the features that best distinguish the subclasses (the Level 2 classes) within the chosen Level 1 class. The two-level approach was motivated by the hierarchical similarity structure of our lung database. Retrieval of images from the same disease class as the query image was not sufficiently accurate; the retrieved images should belong to the same subclass according to image structure, disease stage, and/or severity. Moreover, we observed that the image features that work best to discriminate among different classes are different from the features needed to retrieve "similar" images (i.e., images belonging to the same subclass) within each class. Thus, we customize our features accordingly. The Level 2 step in CQA required that we solve the feature selection for unsupervised data problem (our Level 2 subclasses were unlabeled). We presented our method for performing feature selection and clustering simultaneously (FSSEM). FSSEM wraps feature subset search around EM clustering. EM clustering applies the expectation-maximization algorithm to approximate the maximum likelihood parameter estimate of a finite multivariate Gaussian mixture. Our first experiment on HRCT images of the lungs shows that CQA yields 73.18 percent effective retrieval precision, whereas the single feature vector approach yields only 38.89 percent retrieval precision. Moreover, an evaluation trial with eight radiologists showed that our system, using CQA retrieval, increased our doctors' diagnostic accuracy from 30.2 percent to 63.4 percent.

In the future, it would be interesting to investigate applying CQA on other domains and on extending CQA to more than two hierarchical levels. Currently, CQA returns images from a single major class. Future research would extend CQA to return images from the next most probable major class allowing our system to provide second and third guesses as desired by the user. This would address the central limitation of CQA: the dependency of CQA's retrieved results on the accuracy of the Level 1 classifier.

Acknowledgments

The authors wish to thank the rest of the CBIR group for the discussions and for making our CBIR system possible: Akio Kosaka, Chi-Ren Shyu, Christina Pavlopoulou, Mark Flick, Sean Macarthur, and Alan Marchiori. This research was supported by the US National Science Foundation Grant No. IRI9711535, and NIH Grant No. 1 R01 LM06543-01A1.

References

[1] H. Almuallim and T. Dietterich, "Learning With Many Irrelevant Features," *Proc. Ninth Nat'l Conf. Artificial Intelligence,* AAAI Press, pp. 547-552, 1991.
[2] C. Cardie, "Using Decision Trees to Improve Case-Based Learning," *Machine Learning: Proc. 10th Int'l Conf.,* Morgan Kaufmann, pp. 25-32, 1993.
[3] J. Chen, C.A. Bouman, and J.C. Dalton, "Similarity Pyramids for Browsing and Organization of Large Image Databases," *Proc. SPIE/IS&T Conf. Human Vision and Electronic Imaging III,* vol. 3299, Jan. 1998.
[4] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *J. Royal Statistical Soc., Series B,* vol. 39, no. 1, pp. 1-38, 1977.
[5] M. Devaney and A. Ram, "Efficient Feature Selection in Conceptual Clustering," *Proc. 14th Int'l Conf. Machine Learning,* Morgan Kaufmann, pp. 92-97, 1997.
[6] J.G. Dy, "Feature Selection for Unsupervised Learning Applied to Content-Based Image Retrieval." PhD thesis, Purdue Univ., West Lafayette, IN, 2001.
[7] J.G. Dy and C.E. Brodley, "Feature Subset Selection and Order Identification for Unsupervised Learning," *Proc. 17th Int'l Conf. Machine Learning,* Morgan Kaufmann, pp. 247-254, 2000.
[8] J.G. Dy, C.E. Brodley, A. Kak, C.R. Shyu, and L.S. Broderick, "The Customized-Queries Approach to CBIR Using EM," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 400-406, June 1999.
[9] J.G. Dy, C.E. Brodley, A. Kak, C.R. Shyu, and L.S. Broderick, "The Customized-Queries Approach to CBIR," *SPIE Storage and Retrieval for Image and Video Databases VII,* vol. 3656, pp. 22-32, Jan. 1999.
[10] U. Fayyad, C. Reina, and P.S. Bradley, "Initialization of Iterative Refinement Clustering Algorithms," *Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining,* AAAI Press, pp. 194-198, Aug. 1998.
[11] D.H. Fisher, "Knowledge Acquisition Via Incremental Conceptual Clustering," *Machine Learning,* vol. 2, no. 2, pp. 139-172, 1987.
[12] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, B. Dom, Q. Huang, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The QBIC System," *Computer,* vol. 28, no. 9, pp. 23-32, 1995.
[13] Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Computer and System Sciences,* vol. 55, no. 1, pp. 119-139, 1997.
[14] K. Fukunaga, *Statistical Pattern Recognition,* San Diego: Academic Press, second ed., 1990.
[15] J. Kittler, "Feature Set Search Algorithms," *Pattern Recognition and Signal Processing,* pp. 41-60, 1978.
[16] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence,* vol. 97, no. 1-2, pp. 273-324, 1997.
[17] W.Y. Ma and B.S. Manjunath, "Texture Features and Learning Similarity," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 425-430, 1996.
[18] G.J. McLachlan and K.E. Basford, *Mixture Models, Inference, and Applications to Clustering.* New York: Marcel Dekker, 1988.
[19] T.P. Minka and R.W. Picard, "Interactive Learning Using a 'Society of Models'," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 447-452, 1996.
[20] A. Pentland, R.W. Picard, and S. Sclaroff, "Photobook: Tools for Content-Based Manipulation of Image Databases," *SPIE Storage and Retrieval for Image and Video Databases II,* no. 2185, Feb. 1994.
[21] J.R. Quinlan, "Bagging, Boosting and C4.5," *Proc. 13th Nat'l Conf. Artificial Intelligence,* AAAI Press, pp. 725-730, 1996.
[22] J. Rissanen, "A Universal Prior for Integers and Estimation by Minimum Description Length," *Annals of Statistics,* vol. 11, no. 2, pp. 416-431, 1983.
[23] Y. Rui and T. Huang, "Optimizing Learning in Image Retrieval," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 236-243, June 2000.
[24] S.J. Russell and P. Norvig, *Artificial Intelligence a Modern Approach.* N.J.: Prentice Hall, 1995.
[25] R. Schapire, Y. Freund, P. Bartlett, and W. Lee, "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *Proc. 14th Int'l Conf. Machine Learning,* Morgan Kaufmann, pp. 322-330, 1997.
[26] C. Shyu, C. Brodley, A. Kak, A. Kosaka, A. Aisen, and L. Broderick, "ASSERT: A Physician-in-the-Loop Content-Based Image Retrieval System for HRCT Image Databases," *Computer Vision and Image Understanding,* vol. 75, no. 1-2, pp. 111-132, 1999.
[27] C.-R. Shyu, A. Kak, C.E. Brodley, and L.S. Broderick, Testing for Human Perceptual Categories in a Physician-in-the-Loop CBIR System for Medical Imagery, *Proc. Workshop Content-Based Access of Image and Video Libraries,* pp. 102-108, June 1999.
[28] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1349-1380, Dec. 2000.
[29] P. Smyth, "Clustering Using Monte Carlo Cross-Validation," *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining,* AAAI Press, E. Simoudis, J. Han, and U. Fayyad, eds., pp. 126-133, 1996.
[30] L. Talavera, "Feature Selection as a Preprocessing Step for Hierarchical Clustering," *Proc. 16th Int'l Conf. Machine Learning,* Morgan Kaufmann, pp. 389-397, 1999.
[31] L. Taycher, M. La Cascia, and S. Sclaroff, "Image Digestion and Relevance Feedback in the Image Rover WWW Search Engine," *Proc. Int'l Conf. Visual Information,* Dec. 1997.
[32] K. Tieu and P. Viola, "Boosting Image Retrieval," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 228-235, June 2000.
[33] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions.* John Wiley and Sons, 1985.
[34] S. Vaithyanathan and B. Dom, "Model Selection in Unsupervised Learning with Applications to Document Clustering," *Proc. 16th Int'l Conf. Machine Learning,* Morgan Kaufmann, pp. 433-443, 1999.
[35] W.R. Webb, N.L. Muller, and D.P. Naidich, *High-Resolution CT of the Lung.* PA: Lippincott Williams and Wilkins, third ed., 2001.
[36] J.H. Wolfe, "Pattern Clustering by Multivariate Mixture Analysis," *Multivariate Behavioral Research,* vol. 5, no. 3, pp. 101-116, 1970.