# Content-Based Retrieval from Medical Image Databases: A Synergy of Human Interaction, Machine Learning and Computer Vision

**C. Brodley, A. Kak, C. Shyu, J. Dy**
School of Electrical and
Computer Engineering
Purdue University
West Lafayette, IN 47907
brodley,kak,chiren,dy@ecn.purdue.edu

**L. Broderick**
Department of Radiology
University of Wisconsin Hospital
Madison, WI 53792
lsbroderick@facstaff.wisc.edu

**A. M. Aisen**
Department of Radiology
Indiana University Medical Center
Indianapolis, IN 46202
aaisen@iupui.edu

**Abstract**

Content-based image retrieval (CBIR) refers to the ability to retrieve images on the basis of image content. Given a query image, the goal of a CBIR system is to search the database and return the n most visually similar images to the query image. In this paper, we describe an approach to CBIR for medical databases that relies on human input, machine learning and computer vision. Specifically, we apply expert-level human interaction for solving that aspect of the problem which cannot yet be automated, we use computer vision for only those aspects of the problem to which it lends itself best – image characterization – and we employ machine learning algorithms to allow the system to be adapted to new clinical domains. We present empirical results for the domain of high resolution computed tomography (HRCT) of the lung. Our results illustrate the efficacy of a human-in-the-loop approach to image characterization and the ability of our approach to adapt the retrieval process to a particular clinical domain through the application of machine learning algorithms.

## 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 (Salton, 1986) of the images. Given a query image, the goal of a CBIR system is to search the database and return the $n$ most visually similar images to the query image. A key element of this approach revolves around the types of patterns that can be recognized by the computer and that can serve as the indices of the image retrieval algorithm. Our research addresses the design and implementation of a CBIR system for medical image databases. The success of such an approach provides a unique opportunity to aid physicians in the process of diagnosis.

In the past decade, the field of diagnostic medical imaging has experienced rapid growth and change through both the introduction of new imaging modalities and enhancement in the capabilities of existing techniques. The shift in technology from analog film based methodologies to computer based digital technologies is creating large digital image repositories. CBIR provides an opportunity to tap the expertise contained in these databases in the following way: observing an abnormality in a diagnostic image, the physician can query a database of known cases to retrieve images (and associated textual information) that contain regions with features similar to what is in the image of interest. With the knowledge of disease entities that match features of the selected region, the physician can be more confident of the diagnosis. In our approach to CBIR, an expert radiologist in each domain (anatomic region) selects images for the database, provides the differential diagnosis and when available, includes treatment information. As such, a less experienced practitioner can benefit from this expertise in that the retrieved images, if visually similar, provide the role of an expert consultant.

In this paper, we describe an approach to CBIR for medical databases that relies on human input, machine learning and computer vision. Fundamental to our approach is how images are characterized (indexed) such that the retrieval procedure can retrieve visually similar images within the domain of interest. To aid in the process of adding a new clinical domain, which we define to be a new image modality and anatomic region, our system adapts its image characterization procedure using machine learning algorithms.

In the remainder of this paper we first describe our human-in-the-loop approach to image characterization for medical images and explain why a totally automated approach is not possible or even desirable. We give an overview of a physician's interaction with the system and present salient aspects of the retrieval process. We then outline the steps taken when adding a new clinical domain to the system. This includes a description of the general purpose low-level image features from which our approach selects a customized set for a particular clinical domain. We present empirical results for the domain of high resolution computed tomography (HRCT) of the lung. Our results illustrate the efficacy of a human-in-the-loop approach to image characterization and the ability of our approach to adapt the

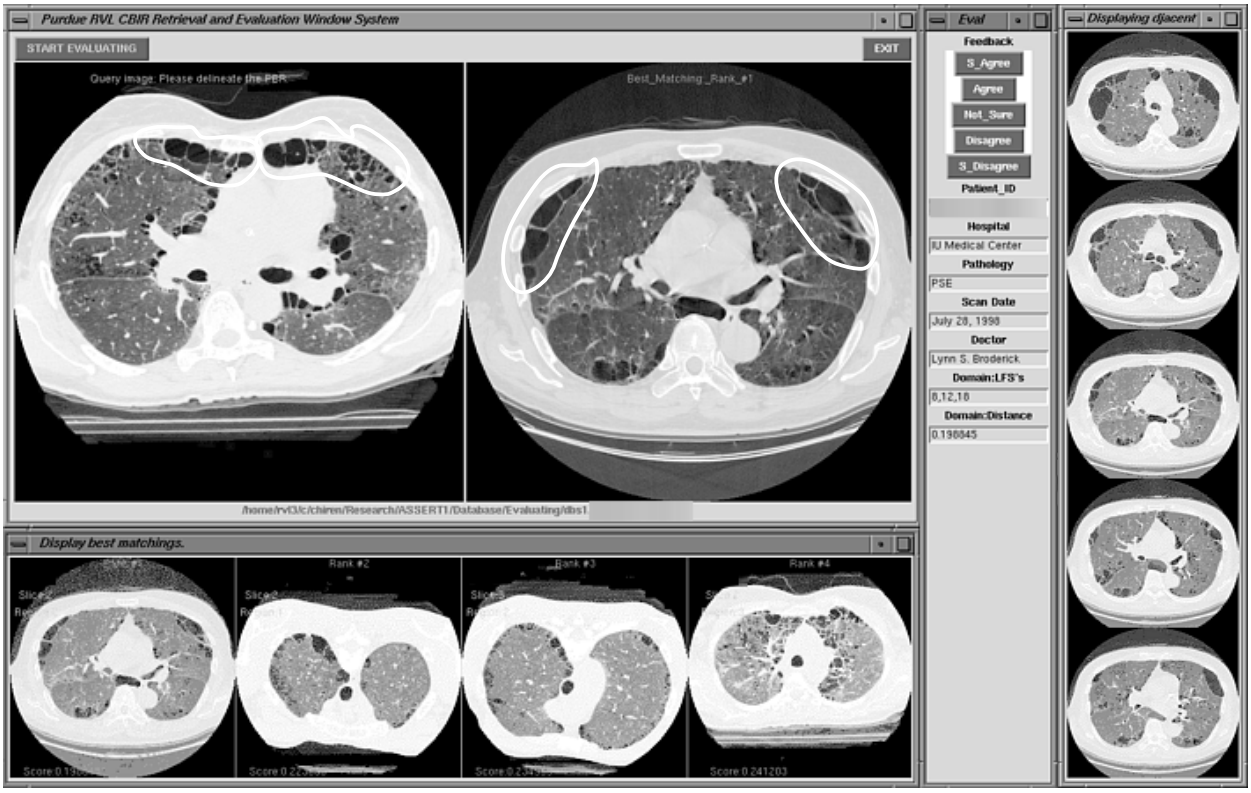



Figure 1: The User Interface

retrieval process to new clinical domains.

## A Physician-in-the-Loop Approach

Given a query image, the goal of a content-based image retrieval system is to return the $n$ most visually similar images to the query image in its database. The most common approach is to characterize the images by a global signature (Flickner, et al, 1995; Kelly, Cannon and Hush, 1995; Stone, and Li, 1996; Pentland, Picard and Sclaroff, 1994; Hou, et al. 1992). For example, the CANDID system (Kelly, Cannon and Hush, 1995) computes histograms from normalized gray levels for image characterization and the QBIC system (Flickner, et al, 1995) characterizes images by global characteristics such as color histogram, texture values and shape parameters of easily segmentable regions.

For medical images, global characterization fails to capture the relevant information (Shyu, et al, to appear). In medical radiology, the clinically useful information consists of gray level variations in highly localized regions of the image. For example, for high-resolution computed tomographic (HRCT) images of the lung, a disease such as emphysema (shown in the circled region in the image in the upper left of Figure 1) manifests itself in the form of a low-attenuation region that is textured differently from the rest of the lung. Attributes characterizing a local region are required because the ratio of pathology bearing pixels to the rest of the image is small, which means that global characteristics such as texture measures cannot capture such local variations.

A human is necessary because the *pathology bearing regions* (PBRs) in our images cannot be segmented out by any of the state-of-the-art segmentation routines due to the fact that for many diseases, these regions often do not possess sharp edges and contours. For example, the PBR's in Figure 2 lack easily discernible boundaries between the pathology bearing pixels and the rest of the lung; however, these PBR's are easily visualized by the trained eye of a physician. Our system, therefore, enlists the help of the physician. Using a graphic interface that we developed, it takes a physician only a few seconds to delineate the PBRs and any relevant anatomical landmarks. A benefit of this approach is that when a query image has more than one pathology, the physician can choose to circumscribe only one of the regions in order to focus retrieval on that pathology.

## A Hierarchical Approach to Image Retrieval

In Figure 1 we show the retrieval results for a query image (shown at left in the main window). The system displays the four best matching images below the main window. For convenience, the user can click on one of these images, causing the system to display a magnified

Figure 2: All three images are from patients with Centrilobular Emphysema

version of the chosen image in the window to the right of the query image. The user can provide feedback in the text window, shown on the right of the enlarged matching retrieved image. Shown in the rightmost column are the additional slices from the patient of the enlarged matching image (for each patient, an HRCT session produces on the order of 20-50 cross-section images, called slices). During image population, our expert radiologist identifies the "key" slices that we then include in the database for indexing and retrieval. Because it can be helpful to view other cross-sections, we retain the extra slices and give the user the ability to browse through them.

Given a query image with an unknown medical diagnosis, we first classify the image as one of the known disease classes. The system then uses the features associated with the predicted class to retrieve the $n$ most similar images, as defined by Euclidean distance, to the query image. This approach is motivated by the observation that the features that are most effective in discriminating among images from different classes may not be the most effective for retrieval of visually similar images within a class. This occurs for domains in which not all pairs of images within one class have equivalent visual similarity – i.e., subclasses exists. For example, the features that we use to distinguish cats from dogs are different than those that we use to distinguish an Australian sheep dog from a collie.

Our approach, which we call *Customized Queries*, is appropriate for many clinical domains, because although each image is labeled with its disease class, within one disease class images can vary greatly with respect to visual similarity on account of the severity of disease and other such factors. Figure 2 illustrates this point. Notice that within the class Centrilobular Emphysema Figure 2c is visually dissimilar to Figures 2a and 2b. Indeed, although a given set of features may be ideal for the disease categorization of a query image, those features may not always retrieve the images that are most similar to the query image. We describe below how machine learning methods are applied to obtain the classifier and the customized feature subsets.

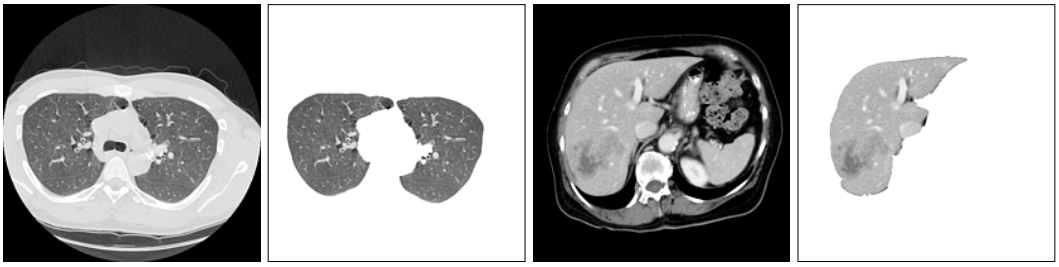
## Handling a New Clinical Domain

Before describing each phase in detail, we give a general overview of the steps needed to add a new clinical domain to our system. The first step is to collect a database of images for which the diagnoses are known. An expert radiologist for that clinical domain provides the images and interacts with our system to delineate all of the PBRs in each image. Currently we are working with experts in the areas of pulmonary lung disease, hepatic disease (liver) and skeletal disease. Once we have collected enough images to make using the system beneficial, we apply our library of computer vision and image processing routines to extract a feature vector of the low level image characteristics for each archived image. At this point we are ready to train the system and to apply machine learning algorithms to build our hierarchical retrieval procedure.

### Image Collection and Region Extraction

We rely on our medical experts to choose representative images. For a given clinical domain, our goal is to ensure a good distribution over the various diseases for two reasons. First, we would like to be able to retrieve at least four images with the same pathology for each query. Second, in order to select the features to use for classification and for retrieval in our customized queries approach we need to obtain sufficient data to make this choice accurately. The ultimate test of whether we have obtained a sufficient number of images in the database is in part measured by the accuracy of our retrieval process. This is best judged by clinicians, and therefore is an inherently subjective measure.

To archive an image into the database, a physician delineates the PBRs and any relevant anatomical landmarks. This interaction takes only a few seconds for a trained domain expert (a radiologist). The left hand image in Figure 1 shows an HRCT image with PBRs as delineated by a physician. The physician also delineates any relevant anatomical landmarks, such as the lung fissures. The information regarding the pathology of the lung resides as much in the location of each PBR with respect to the anatomical markers as it does in the characteristics of the PBRs.



(a) (b) (c) (d)

Figure 3: Region Extraction

The next step is to apply a region extraction algorithm which segments out the tissue type of interest from tissues irrelevant to the disease process. For each new clinical domain, we must write a customized region extraction algorithm. Figure 3a shows the original HRCT of a patient's lungs and 3b depicts the extracted lung region. Figure 3c shows the original CT image of a patient's liver and 3d shows the extracted liver region. Details of these algorithms can be found in (Shyu, et al, to appear).

## General Purpose Image Attributes

To characterize each image, the system computes attributes that are local to the PBRs and attributes that are global to the entire anatomical region.[1] The PBRs are characterized by a set of shape, texture and other gray-level attributes. For characterizing texture within PBRs, we have implemented a statistical approach based on the notion of a gray-level co-occurrence matrix (Haralick and Shapiro, 1992). This matrix represents a spatial distribution of pairs of gray levels and has been shown to be effective for the characterization of random textures. In our implementation, the specific parameters we extract from this matrix are energy, entropy, homogeneity, contrast, correlation, and cluster tendency. In addition to the texture-related attributes, we compute three additional sets of attributes on the pixels within the PBR boundary. The first set computes measures of gray-scale of the PBR, specifically, the mean and standard deviation of the region, a histogram of the local region, and attributes of its shape (longer axis, shorter axis, orientation, shape complexity measurement using both Fourier descriptors and moments). The second set computes the edginess of the PBR using the Sobel edge operator. The extracted edges are used to obtain the distribution of the edges. We compute the ratio of the number of edge pixels to the total number of pixels in the region for different threshold channels, each channel corresponding to a different threshold for edge detection. Finally, to analyze the structure of gray level variations within the PBR, we apply a region-based segmenter (Rahardja and Kosaka, 1996). From the results we compute the number of connected regions in the PBR and histograms of the regions with respect to their area and gray levels.

In addition to the texture and shape attributes, a PBR is also characterized by its average properties, such as gray scale mean and deviation, with respect to the pixels corresponding to the rest of the extracted region. The system also calculates the distance between the centroid of a marked PBR and the nearest relevant anatomical marker (e.g., the lung boundary for the domain of HRCT of the lung). For some domains, we include this anatomical information because physicians use this information to diagnose the patient.

The total number of low-level computer vision attributes is 125. While this gives us an exhaustive characterization of a PBR, for obvious reasons only a small subset of these attributes can be used for database indexing and retrieval. In the next two sections we describe how the retrieval procedure is customized for a given clinical domain.

## Feature Selection for Image Classification

To select the features that will be used to classify a query image (the first level of our customized queries retrieval scheme) our goal is to determine which features provide maximal class separation. The pathology class labels are confirmed diagnoses obtained from medical records, hence we can consider these as ground truth labels.

To find the best classifier, we first extract all 125 features from each database image. We then run a series of experiments using different classifiers coupled with a forward sequential feature selection (SFS) wrapper (Kohavi and John, 1997) using MLC$^{++}$. [2] SFS is a greedy search algorithm that adds one feature at a time. It adds the feature that when combined with the current chosen set of features yields the largest improvement in classification performance. Currently we are favoring forward selection over backward selection as we have found that for a given clinical domain a relatively small set is required. Note that which features are included in

[1] Note that the sense in which we use the word "global" is different from how it is commonly used in the literature on CBIR. Our global attributes are global only to the extent that they are based on all the pixels in the extracted region.

[2] Available at http://www.sgi.com/Technology/mlc

this subset differs from domain to domain. The resulting feature subset and classifier that perform best, as judged by a ten-fold cross-validation over the database, are used to classify the query image during retrieval. Currently we perform feature selection in conjunction with the 1-NN, 5-NN and decision tree algorithms, but there is no reason why other supervised learning algorithms could not be added to the search. Finally, it is important to note that this procedure should be periodically rerun because as we add more images and disease pathologies the set of relevant features and the best classifier may change.

## Feature Selection for Retrieving Visually Similar Images within a Disease Class

After we classify the query image, the next step is to reformulate the query in terms of the feature subset customized for the predicted disease class. In the absence of subclass label information, we must simultaneously find the features that best discriminate the subclasses and at the same time find these subclasses. We resort to unsupervised clustering, which allows us to categorize data based on its structure. The clustering problem is made more difficult when we need to select the best features simultaneously. To find the features that maximize our performance criterion (e.g., retrieval precision), we need the clusters to be defined. Moreover, to perform unsupervised clustering we need the features or the variables that span the space we are trying to cluster. In addition to learning the clusters, we also need to find the optimal number of clusters, $k$. Hence, we have designed an algorithm that for each disease class, simultaneously finds $k$, the clusters and the feature set.

Our approach to feature selection is inspired by the wrapper approach for feature subset selection for supervised learning (Kohavi and John, 1997). 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 the expectation maximization (EM) (Mitchell, 1997; Dempster, Laird, and Rubin, 1977) algorithm and then evaluating the resulting cluster using our chosen clustering criterion. The result of this search is the feature subset that optimizes our criterion function. Because there are $2^n$ feature subsets, where $n$ is the number of available features, exhaustive search is impossible. To search the features, sequential forward, backward elimination or forward-backward search can be used (Fukunaga, 1990). Currently, our system applies sequential forward selection driven by criterion of cluster separability. Because we do not know $k$, the number of clusters, we adaptively search for the value of $k$ during clustering, using Bouman et al's (1998) procedure, which applies a minimum description length penalty criterion to the ML estimates to search for $k$. In the remainder of this section, we provide an overview of our application of the EM algorithm and our chosen separability criterion (full details can be found in (Dy, et al, 1999)).

We treat our data (the image vectors in our database) as a $d$-dimensional random vector and then model its density as a Gaussian mixture of the following form:

$$f(X_i|\Phi) = \sum_{j=1}^{k} \pi_j f_j(X_i|\theta_j)$$

where $f_j(X_i|\theta_j) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_i-\mu_j)^T\Sigma_j^{-1}(X_i-\mu_j)}$, is the probability density function for class $j$, $\theta_j = (\mu_j, \Sigma_j)$ is the set of parameters for the density function $f_j(X_i|\theta_j)$, $\mu_j$ is the mean of class $j$, $\Sigma_j$ is the covariance matrix of class $j$, $\pi_j$ is the mixing proportion of class $j$, $k$ is the number of clusters, $X_i$ is a $d$-dimensional random data vector, $\Phi = (\pi_1, \pi_2, \cdots \pi_k, \theta_1, \theta_2, \cdots \theta_k)$ is the set of all parameters, and $f(X_i|\Phi)$ is the probability density function of our observed data point $X_i$ given the parameters $\Phi$.

The $X_i$'s are the data vectors we are trying to cluster. To compute the maximum likelihood estimate of $f(X_i|\Phi)$ we use the expectation-maximization (EM) algorithm. The missing data for this problem is the knowledge about to which cluster each data point belongs. In the EM algorithm, we start with an initial estimate of our parameters, $\Phi$, and then iterate using the update equations until convergence. The exact form of the update equations can be found in (Dy, et al, 1999).

The EM algorithm can get stuck at a local maxima, hence the initialization values are important. We used $r = 10$ random restarts on $k$-means and pick the run with the highest maximum likelihood to initialize the parameters (Smyth, 1997). We then run EM until convergence (likelihood does not change by more than 0.0001) or up to $n$ iterations whichever comes first for each feature selection search step. (In practice, raising $n$ above 20, does not influence the results). We limit the number of iterations because EM converges only asymptotically, i.e., convergence is very slow when you are near the maximum. Moreover we often do not require many iterations, because initializing with $k$-means starts us at a high point on the hill of the space we are trying to optimize.

Fundamental to any clustering algorithm is the criterion used to evaluate the quality of the clustering assignment of the data points. We applied the $trace(S_w^{-1}S_b)$ criterion (Fukunaga, page 446, 1990). $S_w$ is the within-class scatter matrix and measures how scattered the samples are from their cluster means and $S_b$ is the between class scatter matrix and measures how scattered the cluster means are from the total mean. Ideally, the distance between each pair of samples in a particular cluster should be as close together as possible and the cluster means should be as far apart as possible with respect to the chosen similarity metric. We use the $trace(S_w^{-1}S_b)$ as our criterion because it is invariant under any nonsingular linear transformation, which means that once $m$ features are chosen, any nonsingular linear transformation on these features does not change

the criterion value.

The trace criterion is used to evaluate each candidate feature subset in our feature subset selection search. Note that this procedure selects features that partition the images within a disease class, but that these features do not necessarily correspond to clinically meaningful features. In other work we are investigating whether computer vision methods can capture the perceptual features that physicians say they use to discriminate among different diseases (Shyu, et al, to appear).

## HRCT of the Lung: An Empirical Evaluation of the Approach

Ultimately the true test of a CBIR system is whether it is used by practitioners. To measure whether such a system would be useful, evaluation of an information retrieval system is done by measuring the recall and the precision of the queries. Recall is the proportion of relevant materials retrieved. Precision quantifies the proportion of the retrieved materials that is relevant to the query. In our approach, the precision and recall are functions of 1) the attribute vector used to characterize the images, 2) the delineation of the PBR by the physician, and 3) the retrieval scheme.

The experimental results presented in this section were designed to meet two goals. First to evaluate the contribution made by local characterization, which comes at the price of needing human interaction. Second, to evaluate the ability of the supervised and unsupervised machine learning methods to correctly identify the features used in our hierarchical retrieval scheme. In this paper, we present results using the image modality of high resolution computed tomography images and the clinical domain of pulmonary lung disease.

Our current HRCT lung database consists of 312 HRCT lung images from 62 patients. These images yield 518 PBRs. A single image may have several PBR's and these PBR's may have different diagnoses. Throughout the experiments we considered each PBR as a data point, i.e., a single image with three PBR's gives us three data points. These images were identified by radiologists during routine medical care at Indiana University Medical Center. Currently, the diseases in the database are centrilobular emphysema (CE), paraseptal emphysema (PE), sarcoid (SAR), invasive aspergillosis (ASP), broncheitasis (BR), eosinophilic granuloma(EG), and idiopathic pulmonary fibrosis (IPF). The number of PBRs of each disease is shown in the first column of Table 1.

### Local versus Global Image Characterization

This experiment is designed to test the utility of characterizing medical images using local rather than global attributes. To ensure a situation that would mirror its use in a clinical setting, we omit the query-image patient's images from the database search (each patient may have more than one image in the database to ensure a distribution over the different ways in which the disease can appear in an image). Our statistics were generated from the four highest ranking images returned by the system for each query.

Table 1: Retrieval Accuracy of Global versus Localized Attributes.

| Disease Class | | Correct Retrievals | | Percent of Total | |
|---|---|---|---|---|---|
| | | $F_L + F_C$ | $F_G$ | $F_L + F_C$ | $F_G$ |
| CE | 314 | $2.92 \pm 0.18$ | $2.12 \pm 0.85$ | 73 | 53 |
| PE | 54 | $3.04 \pm 0.27$ | $1.68 \pm 1.07$ | 76 | 42 |
| IPF | 51 | $2.88 \pm 0.32$ | $2.08 \pm 0.14$ | 72 | 52 |
| EG | 57 | $2.72 \pm 0.15$ | $1.92 \pm 0.32$ | 68 | 48 |
| SAR | 16 | $2.76 \pm 0.71$ | $1.96 \pm 0.75$ | 69 | 49 |
| ASP | 12 | $1.92 \pm 0.80$ | $1.64 \pm 0.36$ | 48 | 41 |
| BR | 14 | $3.00 \pm 0.32$ | $2.32 \pm 0.55$ | 75 | 58 |
| Total | 518 | $2.88 \pm 0.23$ | $2.03 \pm 0.72$ | 72 | 51 |

Table 1 shows results for two different sets of attributes. The first is a combination of attributes extracted from the PBR region ($F_L$) and attributes contrasting the PBR to the rest of the lung region ($F_C$). The combined set $F_L + F_C$ was chosen by the SFS algorithm wrapped around a one-nearest neighbor classification algorithm. The second set of attributes $F_G$ was customized to a global approach to image characterization. The $F_G$ attributes were chosen by the SFS algorithm when optimizing performance for the entire lung region. For this experiment we used the nearest-neighbor retrieval method, which retrieves the four images closest to the query image as measured by the Euclidean distance of the chosen features. For each disease category in our database, we show the mean and standard deviation of the number of the four highest ranking images that shared the same diagnoses as the query image, and percentage of the four retrieved images that have the same diagnosis as the query image.

The attributes in $F_L$ are: the gray scale deviation inside the region, gray-level histogram values inside the region, and four texture measurements (homogeneity, contrast, correlation and cluster). The attributes in set $F_C$ contrasting the PBR to the entire lung are: the area of the PBR, the Mahalanobis distance from the centroid of PBR to the nearest lung boundary point, the difference of gray-scale mean of the PBR and the entire lung, and the difference of gray-scale deviation of the PBR and the entire lung. The attributes in set $F_G$ are: gray scale mean and deviation, histogram distribution, histogram distribution after gamma correction, and four texture measures (cluster, contrast after gamma, cluster after gamma, and edginess of strength after gamma). From the table we see that the localized image characterization method ($F_L + F_C$) has higher precision than the global image characterization method, illustrating that local attributes significantly improve retrieval performance in the domain of HRCT of the lung.

Table 2: Retrieval Results for the Domain of HRCT of the Lung.

| Disease Class | Number of Queries | $k$ | Traditional Method | | | | | Customized Queries | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | SA | A | NS | D | SD | SA | A | NS | D | SD |
| CE | 18 | 5 | 28 | 9 | 5 | 2 | 28 | 69 | 2 | 1 | 0 | 0 |
| PE | 3 | 4 | 0 | 0 | 4 | 0 | 8 | 10 | 0 | 1 | 0 | 1 |
| IPF | 2 | 3 | 5 | 0 | 0 | 0 | 3 | 3 | 2 | 2 | 0 | 1 |
| EG | 1 | 4 | 0 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 0 |
| SAR | 1 | 5 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 4 |
| ASP | 1 | 5 | 0 | 0 | 0 | 0 | 4 | 3 | 1 | 0 | 0 | 0 |
| BR | 1 | 2 | 0 | 0 | 0 | 0 | 4 | 3 | 1 | 0 | 0 | 0 |
| total | 27 | | 33 | 9 | 9 | 2 | 55 | 92 | 6 | 4 | 0 | 6 |

One concern of a physician-in-the-loop approach is that precision is a function of PBR delineation. To address this concern, we have performed a sensitivity analysis of our ability to classify PBR to physician subjectivity. Using the same experimental setup, we compared the retrieval results of the physician marked PBRs to larger and smaller PBRs. An empirical analysis illustrated that shrinking or growing the PBR by 50% had a less than 3% impact on the classification accuracy of our method.

### The Traditional Approach versus Customized Queries

This experiment illustrates that customized queries[3] results in better retrieval precision than the traditional approach to CBIR, which retrieves the $n$ closest images in the database as measured using the Euclidean distance of the features selected to optimize the accuracy of a 1-NN classifier. In assessing the performance of customized queries we assumed an 100% accurate classifier was used to classify a query as its disease class. We did this to isolate the effect of using the appropriate customized features in retrieving the images, i.e., the utility of customizing a query. This assumption is not too limiting since the classification accuracy we obtained from a ten-fold cross-validation applied to a 1-NN classifier of the disease classes is $93.33\% \pm 0.70\%$.[4] In our conclusions we address what steps we take when unacceptable retrieval results are obtained due to an inability to classify the image correctly.

To determine which method is best, the lung specialist in our team was asked to evaluate the retrieval results of the two approaches. Throughout the test, the radiologist was not informed as to which method produced the retrieved images.[5] In Table 2 we show the number of queries evaluated for each disease. We chose eighteen from C-Emphysema because it is the largest class in our collection (51% of our database is of class C-Emphysema). The number of clusters chosen for each class is shown in column 2. 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. To evaluate the system, the user 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 SA, 1 for A, 0 for NS, $-1$ for D and $-2$ for SD.

The traditional approach received a total of $-37$ points, whereas customized queries received 178 points. If SA and A are considered as positive retrievals and the rest as negative retrievals, The traditional approach resulted in 38.89% retrieval precision and customized queries resulted in 90.74% precision. Notice that for the traditional approach precision is not the same as the accuracy obtained for the disease class classifier because there were cases for which the radiologist did not mark SA or A even though the retrieved images had the same diagnosis as the query image. From these results, we can see that customized queries dramatically improves retrieval precision compared to the traditional approach for this domain.

## Conclusions and Future Work

In this paper we presented our approach to content-based image retrieval for medical images, which combines the expertise of a human, the image characterization from computer vision and image processing, and the automation made possible by machine learning techniques. We believe that our system combines the best of what can be gleaned from a physician, without burdening him or her unduly, and what can be accomplished by computer vision and machine learning.

In an empirical evaluation, we demonstrated that local attributes significantly improve retrieval performance in the domain of HRCT images of the lung over a purely global approach. A sensitivity study showed that physician subjectivity in PBR delineation impacts

[3]Note that for the results presented here, $k$ was chosen using a separability criterion (Dy, et al, 1999).

[4]Note that the retrieval precision in Table 1 was not 90% because in the table we are measuring how many of the four nearest neighbors have the same disease label, whereas here the 93.33% reports the percentage of time that the nearest neighbor has the same disease label as the query image.

[5]To keep the radiologist from guessing, we randomly interleaved the two methods.

performance by only a negligible amount. In a clinical trial, we illustrated that customized queries significantly improve retrieval prevision over the traditional single vector approach to CBIR as evaluated on the domain of HRCT of the lung.

We are working on several fronts to improve our approach. In addition to the domain of HRCT of the lung, we are in the process of populating databases in the domains of CT of the liver, MRI of the knee and MRI of the brain. One potential drawback of the customized queries approach is that when the classifier misclassifies a query image, the retrieval procedure customizes the query to the wrong class. To mitigate this effect, when a physician does not enter agree (or strongly agree) for at least two of the retrieved images, we try again by resorting to the traditional method of retrieval which searches the entire database. Furthermore, although we have an overall classification accuracy of approximately 93%, this accuracy is not uniform across disease classes. For less populous disease classes, the accuracy can be far lower. One reason for this is that the supervised feature selection process does not take this uneven distribution into account. We are currently working on how to select features such that classification accuracy on the less populous classes is not sacrificed for the dominant classes. To help in this endeavor we will investigate how to combine other text-based information about the patient to aid in the initial classification of the pathology bearing region in the image, such as the results of blood tests, age, etc. Finally, we are investigating how to best use user feedback when retrieval results are judged unsatisfactory.

## Acknowledgments

This work is supported by National Science Foundation under Grant No. IRI9711535 and the National Institute of Health under Grant No. 1 R01 LM06543-01A1. We would like to thank Mark Flick and Sean MacArthur for their ideas and work on system design.

## References

Bouman, C., Shapiro, M. Cook, G., Atkins, C. and Cheng, H. 1998. CLUSTER: An unsupervised algorithm for modeling Gaussian mixtures, http://dynamo.ecn.purdue.edu/~bouman/software/cluster

Dempster, A., Laird, N. and Rubin, D. 1977, Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Statistical Society*, B, vol. 39, no. 1, pp. 1-38.

Dy, J. G., Brodley, C. E., Kak, A. Shyu, C. and Broderick, L.S., 1999. The customized-queries approach to CBIR Using EM, *Computer Vision and Pattern Recognition*, Fort Collins, CO, June 1999.

M. Flickner, et al, 1995 Query by image and video content: The QBIC system, *IEEE Computer*, pp. 23-32, September 1995.

Fukunaga, K. 1990, *Introduction to Statistical Pattern Recognition*, 2nd Edition, Academic Press.

Haralick, R. M. and Shapiro, L. G., 1992. *Computer and Robot Vision*, Addison-Wesley.

Hou, T. Y., Hsu, A., Liu, P., and Chiu, M. Y., 1992. A content-based indexing technique using relative geometry features, *Proc. SPIE Conf. on Storage and Retrieval for Image and Video Databases,* pp. 29-98.

Kohavi, R. and John, G. 1997. Wrappers for feature subset selection, *Artificial Intelligence Journal*, Vol. 97, No.s 1-2, pp. 273-324.

Kelly, P. M., Cannon, T. M., and Hush, D.R., 1995. Query by image example: The CANDID approach, in *SPIE Vol. 2420 Storage and Retrieval for Image and Video Databases III*, pp. 238-248.

Mitchell, T. 1997. *Machine Learning*, pp. 191-196, McGraw-Hill Companies Inc.

Pentland, A., Picard, R. and Sclaroff, S. 1994. Photobook: Tools for content-based manipulation of image databases, *Proc. SPIE Conf. on Storage and Retrieval for Image and Video Databases,* pp. 34-47.

Rahardja, K. and Kosaka, A. 1996. Vision-based bin-picking: Recognition and localization of multiple complex objects using simple visual cues, in *1996 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Osaka, Japan, November, 1996.

Salton, G. 1986. Another look at automatic text-retrieval systems. *Communications of the ACM*, 29(7), pp 648-656.

Shyu, C., Brodley, C., Kak, A., Kosaka, A., Aisen, A. and Broderick, L., to appear. ASSERT, a physician in the loop content-based image retrieval system for HRCT image databases, *Computer Vision and Image Understanding.*

Smyth, P. 1996. Clustering using Monte Carlo cross-validation, *The Second International Conference on Knowledge Discovery and Data Mining*, pp 126-133.

Stone, H. S. and Li, C. S., 1996. Image matching by means of intensity and texture matching in the Fourier domain, *Proc. SPIE Conf. in Image and Video Databases,* San Jose, CA, Jan. 1996.

Wolfe, H. H. 1970. Pattern clustering by multivariate mixture analysis, *Multivariate Behavioral Research*, vol. 5, no. 3, pp 101-116.