This paper appears in: Proceedings of the IAPR Workshop on Machine Vision Applications, 2000

Computer Vision Techniques for Content-Based Image Retrieval from Large Medical Databases

Avi Kak and Christina Pavlopoulou *
School of Electrical and Computer Engineering
Purdue University

Abstract

The main goal of content based image retrieval is to efficiently retrieve images that are visually similar to a query image. In this paper we will focus on content based image retrieval from large medical databases, outline the problems specific in this area and describe the recent advances in the field. We will also present some of the more significant results obtained with ASSERT (Automatic Search and Selection Engine with Retrieval Tools), the content based image retrieval system developed in our laboratory.

1 Introduction

Content Based Image Retrieval (CBIR) has emerged during the last several years as a powerful tool to efficiently retrieve images visually similar to a query image. The main idea is to represent each image as a feature vector and to measure the similarity between images with distance between their corresponding feature vectors according to some metric. Finding the correct features to represent images with, as well as the similarity metric depend on the image domain and the goal of the retrieval system.

The list of CBIR systems developed today is long and includes but is not limited to QBIC [24], CANDID [5], PhotoBook [17], MARS [14], ImageRover [7]. Usually these systems are founded on the premise that images can be characterized by global signatures. For example, the CANDID system [5] computes histograms from normalized gray levels for image characterization and the QBIC system [24] characterizes images by global characteristics such as color histogram, texture values and shape parameters of easily segmentable regions.

Medical CBIR systems differ from general purpose CBIR systems in very significant ways. First of all, it is often the case that the clinically useful information consists of gray level variations in highly localized regions of the image. For example, in a high-resolution computed tomographic (HRCT) image of the lung, a disease such as emphysema (shown in Figure 1) manifests itself in the form of low-attenuation regions that are textured differently from the rest of the lung. Local attributes are needed for such situations because the number of pathology bearing pixels in an image is small relative to the number of pixels in the rest of the image, and any global signature would not be sufficiently impacted to serve as a useful attribute for image retrieval. This bodes ill for many of the previously developed methods for CBIR with regard to their use for medical radiology.

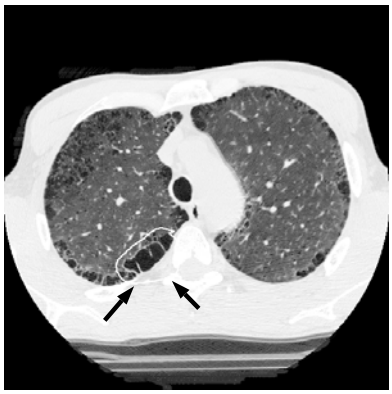


Figure 1: *HRCT lung image with a pathology bearing region delineated by an expert physician (white contour pointed to by dark arrows).*

The need to characterize localized areas of an image induces an additional complication in medical image retrieval systems: such regions may or may not be automatically segmentable from the image. When the shapes of these local regions are highly distinct, as with the ventricular regions in the MR scans reported in [6] and high-contrast single tumor regions in the images of [18], automatic segmentation techniques may succeed. But obviously for the kinds of images shown in Fig. 1, there is no chance that any of the automatic segmentation techniques

*Address: School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, U.S.A. E-mail: {kak,pavlo}@ecn.purdue.edu

known today would work. It is for this reason that in ASSERT [2] we enlist the help of a physician for delineating the pathology bearing regions (PBR) and any other relevant anatomical landmarks.

The above discussion motivates us to define three areas which are important in the creation of a successful medical CBIR system. First, what features are needed to represent the useful information in the image? Often the features needed are too many, so it is necessary to employ a feature selection algorithm to reduce the dimension of the space. Second, what retrieval algorithm should one use, so that similar images are the ones that belong to the same disease class? Third, how could one use user input to improve the retrieval result? This is often referred to in the CBIR literature as relevance feedback. Since we will use ASSERT to illustrate these issues, we will start with a very brief description of the overall system in the next section.

2 ASSERT Architecture

Figure 2 illustrates the different modules of our system for content-based image retrieval from a database of HRCT images. To apply ASSERT to a different domain only the shaded modules would need be replaced. Initially, a physician delineates the PBRs and any relevant anatomical landmarks. The system then executes a suite of image processing algorithms to create the feature vectors that characterize the PBRs individually.

We have experimented with two different sets of features. The first set contains a large number of "generic" low-level features that capture texture and shape information. They are "generic" in the sense that they do not require domain knowledge to be computed, and they could be applied to other imaging modalities without significant changes. The second set aims at capturing the visual cues (or perceptual categories) used by physicians to make a diagnosis. The rationale behind that is that it leads to a more disciplined way of determining what low-level features to extract. It assumes of course that the domain expert knowledge can be elicited, which is true for the HRCT domain.

Subsequently, a feature selection procedure is applied to reduce the dimensionality of the feature space. In figure 2, a decision tree based approached is used as the default retrieval method, however we have experimented with other approaches as well. For more information about the other system modules the reader is referred to [2].

3 General Purpose Image Features

With regard to general purpose features we compute features that are local to the PBRs and features that are global to the entire lung region. [1] The PBRs are characterized by a set of shape, texture and other gray-level attributes compute with standard image processing techniques [10, 21]. For characterizing texture within PBRs, a statistical approach based on the notion of a gray-level co-occurrence matrix has been implemented. This matrix represents a spatial distribution of pairs of gray levels and has been shown to be effective for the characterization of random textures. The specific parameters we extract from this matrix are energy, entropy, homogeneity, contrast, correlation, and cluster tendency. In addition to the texture-related features, we compute three additional sets of features on the pixels within the PBR boundary. The first set computes measures of gray-scale of the pathology bearing region, 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. The ratio of the number of edge pixels to the total number of pixels in the region is computed 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, a region-based segmenter is applied. From the results the number of segmented regions per area and histograms of the area and gray-levels of the segmented regions are computed.

In addition to the texture and shape features, 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 lung. Measurement of these properties requires that we be able to segment out the lung region (note that the lung region is also needed for the measurement of the global features we mentioned earlier). To extract the lung region, we apply a set of binary-image analysis routines [2].

The total number of features, 255 in number, computed for a PBR is large (details of the set of features can be found in [2]). While this gives us an exhaustive characterization of a PBR (an intentional aspect of our design), for obvious reasons only a small subset of these features can be used for database indexing and retrieval.

The features actually used are found by applying the Sequential Forward Search (SFS) algorithm ([12, 13]) to all the 255 features. SFS is a greedy al-

[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 features are global only to the extent that they are based on all the pixels in the entire lung region.

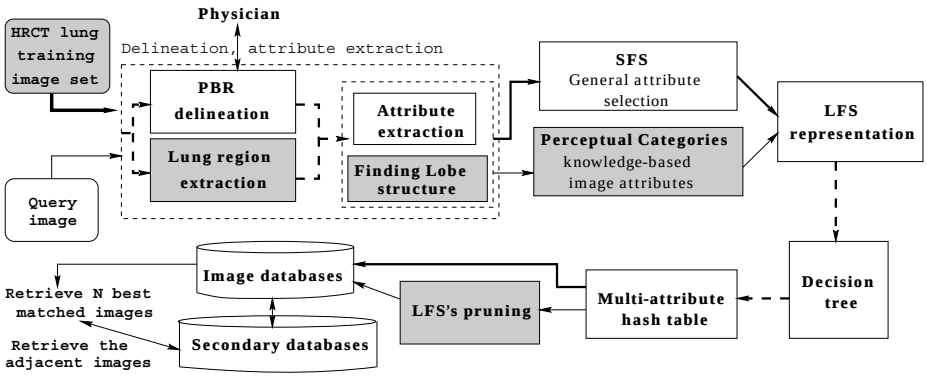


Figure 2: *ASSERT Architecture*

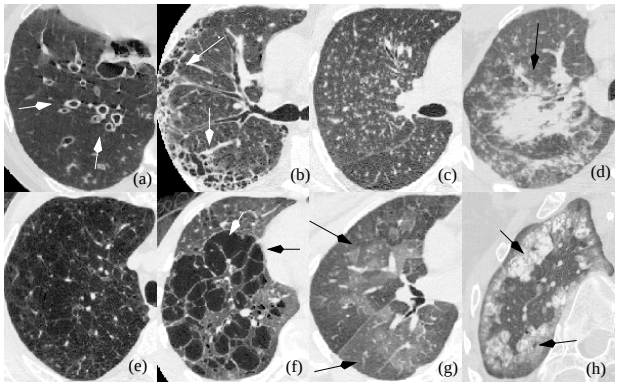


Figure 4: *Perceptual categories used by expert physicians: (a) Bronchial structure (linear & reticular), (b) Honeycombing (linear & reticular), (c) small nodules (nodular opacities), (d) big nodules (nodular opacities), (e) low-attenuation (low opacities), (f) cystic structure (low opacities), (g) ground-glass (high opacities), (h) calcification (high opacities).*

gorithm 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.

4 Features Derived from Physicians' Perceptual Categories

Perceptual category features are motivated by the visual structures the physicians use for diagnosing diseases in HRCT images of the lung (Fig. 3). The four major categories are [15]: *linear and reticular opacities*, *nodular opacities*, *diffuse regions of high attenuation*, and *diffuse regions of low attenuation*. These categories can be called major in the sense that, in the physician's mind, they possess a strong one-to-one correlation with the various lung diseases. The leaf nodes of the tree in Fig. 3 show the subcategories that the physicians actually use for labeling the PBRs. A PBR may exhibit a pathology corresponding to the major category "Linear & Reticular", but the actual visual structure inside the PBR would either be linear or reticular, corresponding to the two leaf nodes in Fig. 3.

4.1 Linear and reticular opacities

As the name implies, these patterns consist of line-like structures that can either be straight and elongated, web-like, or circular with a dot-like protrusion (the last is also referred to as a signet-ring pattern). These visual structures are most often a result of the thickening of the walls of the bronchi (Fig. 4(a)) and peripheral honeycombing (Fig. 4(b)). Since the walls of the bronchi are characterized by adjacent low and high attenuation regions, they can be extracted by dual-thresholding [21]. The following low-level features measure the relevant characteristics of such structures: *the number of bronchial objects* and *the average thickness of the bronchi-walls*. Reticular patterns that show up as peripheral honeycombing respond to the skeletonization of the PBR, followed by the extraction of the following parameters: *the number of cells formed by the skeleton*, *the average cell size*, and *the number of cells adjacent to the lung boundaries or fissures*.

4.2 Nodular opacities

The gray values associated with nodular opacities carry important information with regard to whether the tissue is benign or malignant. HRCT images that show this type of evidence can be further categorized on the basis of the size and locational distributions associated with the nodular opacities. The nodular opacities appear typically in two different sizes: small nodules, which are roughly round and less than one centimeter in diameter, and large nodules of irregular shape, whose "diameter" exceeds one centimeter. Sometimes large nodules agglomerate into large masses, as shown in Fig. 4(d). For the case of small nodules, their distribution carries

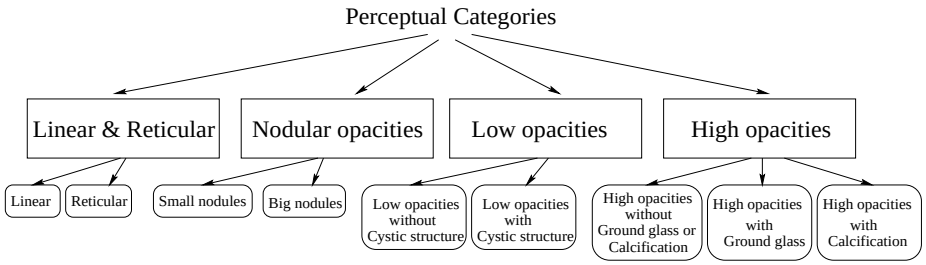


Figure 3: *The perceptual category tree.*

diagnostic information. When the distribution is random, then the nodules appear widely and evenly throughout the lung as shown in Fig. 4(c). Distributions become non-uniform when nodules attach themselves to the boundaries of the lungs or to the fissures. Images with nodules respond to feature extraction algorithms in which the system first applies a threshold to the lung regions, followed by the measurement of "roundness" property. The roundness property is particularly effective for extracting small nodules. The large nodules are extracted with a lower threshold on the roundness parameter. In other words, the value of the roundness threshold is keyed to the size of the object extracted after thresholding. Effective feature measurements for images with this type of pathology include *the average sizes of nodules*, *average roundness of nodules*, *average nearest-neighbor distance between the nodule centers*, and the *gray-level mean of nodules*.

4.3 Diffuse regions of high attenuation (high opacities)

For some of the lung diseases, the entire lung may assume a different shade of gray in comparison to a normal lung. For example, shown in Fig. 4(g) is what is referred to as *ground-glass opacity*. When present, it does not obscure the underlying vessels, that is the vessels can be seen clearly in the lungs even though the tissues everywhere are characterized by a higher level of attenuation. Lumped in the same perceptual category is the pattern that corresponds to *calcification* shown in Fig. 4(h). The overall visual effect gleaned from the HRCT image is that of marked increase in density, similar to bone. Algorithms capable of separating the normal tissues from the ground-glass tissues make use of the fact that gray-level histogram for the latter case is strongly bimodal, whereas it is primarily unimodal for the normal tissues. After the ground-glass tissues are extracted, the vascular structure is extracted by employing the well-known technique of co-occurrence matrices [10] with different values for the orientation parameter. The computed measurements are *uniformity of energy*, *homogeneity*, *gray level mean of ground-glass regions*, and *the ratio of abnormal regions and lung regions*.

4.4 Diffuse regions of low attenuation (low opacities)

All of the previously mentioned perceptual categories are marked by increased attenuation (meaning higher gray levels) associated with the pixels corresponding to the diseased tissues. The category we will describe in this section is marked by *decreased* attenuation. For example, *centrilobular emphysema* shows up in HRCT images in the form of a large number of areas with markedly decreased density, as shown in Fig. 4(e). These areas may occupy the entire lung region, but are likely to predominate in the upper lobes. When the disease becomes severe, these areas may join together to form a large region of low attenuation. This perceptual category also includes low-attenuation blobs bounded by a high-attenuation background (Fig. 4(f)). These visual structures respond to the following feature extraction steps: First, the normal tissues and the low-attenuation tissues are separated by simple thresholding. (The gray level histogram is strongly bimodal for all these diseases.) Next, the co-occurrence matrices are computed for the low pixels resulting from thresholding. Additionally, *the number of decreased density regions adjacent to the lung boundaries or fissures* is also computed, as it carries diagnostic information for the diseases mentioned in this section.

4.5 Are The Low-Level Features Measuring The Physicians' Perceptual Categories?

We have used multivariate analysis of variance (MANOVA) [11, 3] to determine whether or not the low-level features we use for determining the presence or the absence of the perceptual categories are doing their job. MANOVA is used to compute the means of the low-level features separately for the different perceptual categories; the between-category differences of these means; and a measure of the

power of the low-level features to discriminate between the different perceptual categories.

The PBRs labeled by a physician are grouped into nine perceptual categories, corresponding to the leaves of the tree shown in Fig. 3. We shall use the following symbols to refer to these nine categories: linear (G_{linear}), reticular ($G_{reticular}$), small nodules (G_{s_nodule}), big nodules (G_{b_nodule}), high-opacities (G_{high}), low-opacities (G_{low}), cystic structure (G_{cystic}), ground-glass (G_{gg}), and calcification (G_{cal}). To keep the MANOVA part of the discussion general, we will use N_c to denote the number of perceptual categories.

For the purpose of applying the tools of MANOVA, each observation consists of a vector of p low-level feature measurements from a PBR. Note that the p low-level features for category A will, in general, be different from the p low-level features for category B. Additionally, the value of p for category A is allowed to be different from the value of p for category B. This point is important because the categories do not reside in the same p-dimensional feature space. A p-dimensional feature vector is used to set a given category apart from all other categories.

Before MANOVA can be applied, the data must satisfy certain assumptions. The most notable of these are: 1) each observation $X_{g,k}$ is a random sample from perceptual category g; 2) the random samples from different categories are independent; and 3) the distribution corresponding to each category is multivariate normal. We believe that our data does indeed satisfy the first two assumptions. With regard to the third assumption, at this time we have taken it as an article of faith, to be tested more rigorously in the months to come. Since tools like Kolmogorov-Smirnoff tests are available for testing this assumption, the reader might wonder why we haven't applied such a test. Currently, the sparseness of the data for some of the perceptual categories precludes such an analysis. But, as we accumulate more data, this problem will disappear.

Although MANOVA could be used to analyze simultaneously the data for all the categories (in order to determine whether or not sufficient discrimination is provided by the features), it is more efficient to proceed in the following manner: Let N_T be the total number of observations available for all perceptual categories and let N_g be the total number of observations, or sample vectors, for category g. We now divide the data into two sets, one consisting of the N_g samples of category g and the other consisting of the remaining $N_T - N_g (= N_{\rm rest})$ samples. For this two-class problem, we can then test the hypothesis that the p features are able to differentiate between category g and the rest of the data. The data set consisting of the $N_{\rm rest}$ samples will be denoted $X_{\rm rest}$ and the mean of this data by $\overline{X_{\rm rest}}$

This hypothesis testing would, of course, need to be carried out separately for each category. For the remaining discussion here, we will use $X_{g,k}$ to denote the $k^{\rm th}$ observation in category g. And while we are analyzing the data for category g, we will use $X_{rest,k}$ to denote the $k^{\rm th}$ sample of the rest of the data. The mean sample vector for category g is denoted $\overline{X_g}$. We will use Σ to denote the covariance matrix of all the N_T samples of data.

In the p-dimensional space used for category g, it is possible to express an observation vector $X_{g,k}$ by:

$$X_{g,k} \;=\; \overline{X} \;+\; \left(\overline{X_g} - \overline{X}\right) \;+\; \left(X_{g,k} - \overline{X_g}\right) \qquad (1)$$

where $\overline{X}$ is the overall sample mean. This decomposition highlights the contribution made by the deviation of the observation vector from its own category mean and the difference between a category mean and the entire population mean. The latter will be denoted by $\tau_g \;=\; \left(\overline{X_g} - \overline{X}\right)$. In the same p-dimensional space, the expression for the overall covariance of the data can now be expressed as:

$$\sum_{i \in \{g, rest\}} \sum_{k=1}^{N_i} (X_{i,k} - \overline{X})(X_{i,k} - \overline{X})^T$$

$$= \sum_{i \in \{g, rest\}} N_i (\overline{X_i} - \overline{X})(\overline{X_i} - \overline{X})^T \;+$$

$$\sum_{i \in \{g, rest\}} \sum_{k=1}^{N_i} (X_{i,k} - \overline{X_i})(X_{i,k} - \overline{X_i})^T \qquad (2)$$

$$\mathbf{T} \;=\; \mathbf{B} + \mathbf{W} \qquad (3)$$

This shows that the overall data variance $\mathbf{T}$ consists of two parts: $\mathbf{B}$: the between category variance, which has $d_B = 1$ degree of freedom for the two-class problem we are analyzing here; and $\mathbf{W}$: the within category residual variance with $d_W = \sum_{i \in \{g, rest\}} N_i - 2$ degrees of freedom.

To determine whether or not there exists category discrimination information in the low-level features used to measure the presence or absence of a category in a PBR, we can perform the following likelihood ratio test. We construct a hypothesis $\mathbf{H_0} : \tau_{\rm g} \;=\; \tau_{\rm rest}$, meaning that the mean for category g is the same as the mean for all other categories lumped together within a chosen confidence interval in the p-dimensional space specific to category g. $\tau_{\rm rest}$ denotes $\overline{X_{\rm rest}} - \overline{X}$. To test the $\mathbf{H_0}$ hypothesis, we first compute *Wilks' lambda* Λ^*:

$$\Lambda^* \;=\; \frac{|\mathbf{W}|}{|\mathbf{B} \;+\; \mathbf{W}|} \qquad (4)$$

The exact distribution of Λ^* can be obtained from any standard published table if the size of the category vector is known. A criterion derived from

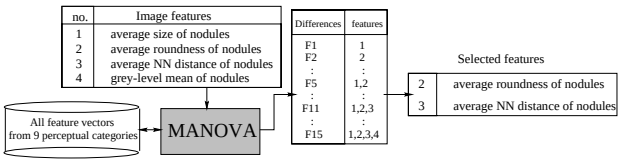


Figure 5: *Choosing the maximally discriminatory features for the small nodule perceptual category.*

the applicable distribution can then be compared against a threshold for either accepting or rejecting the hypothesis H_0 at a chosen confidence level. For example, when each observation vector consists of two low-level features, meaning $p = 2$, the following F-test criterion obtained from the applicable distribution

$$F \;=\; \left(\frac{d_W - 1}{d_B}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \qquad (5)$$

can be compared to a threshold as follows

$$F \;>\; F_{d_B,d_W}(\alpha) \qquad (6)$$

to reject hypothesis H_0 at confidence level $(1-\alpha)$. $F_{d_B,d_W}(\alpha)$ is the upper $100\alpha\%$ of the F-distribution with d_B and d_W degrees of freedom.

In this manner, we can determine whether or not a given p-dimensional feature set can discriminate a category vector from the rest of the data. This pairwise hypothesis testing is carried out separately for all the categories.

4.6 Choosing the Maximally Discriminatory Feature Set

In Section 4, we described the low-level features that could be used for determining the presence or the absence of each of the perceptual categories in an image. For each perceptual category, we test the H_0 hypothesis for all combinations of the low-level features listed in that section at $\alpha = 0.1$ level. For example, for the category $G_{s\_nodules}$ we start with the four features listed in Fig. 5. The H_0 hypothesis is tested for all combinations of these four features. The total number of these feature combinations is $\sum_{i=1}^{4} C_i^4 = 15$. The F value of Section 4.5 was used to determine the quality of each feature combination. We selected the feature combination that corresponds to the highest F value. Fig. 5 shows this process pictorially. In this case, the subset {2,3} produced the highest F value. This process is repeated for each perceptual category.

4.7 Weighting the Low-Level Features

If the inequality of Eq. 6 holds for the aforementioned pairwise hypothesis testing for each of the categories, we can conclude that the chosen low-level features discriminate between the prescribed perceptual categories. This also means the sets of image features are good for classifying PBRs based on the perceptual categories. But the following questions remain: What is the relative contribution of each of the low-level features to the differences in the means of the different categories? Could knowledge of these relative contributions be used to weight the image features differently? This section addresses these two questions.

To assess the relative weights to be assigned to the individual low-level features, we used the Bonferroni method of multiple comparisons. For the sake of explanation, let's assume that we have only three perceptual categories: G_{cystic}, $G_{reticular}$, and G_{s_nodule}. Let the following two low-level features be designated as being capable of discriminating between the category G_{cystic} and the other categories: *number of cells* and *average size of cells*. Let's assume that this feature set rejects the hypothesis H_0 at confidence level $1-\alpha$.

To ascertain the relative importance to be assigned to each G_{cystic} feature, we compute the differences in the means of the feature values for the following pairs of categories: (G_{cystic}, $G_{reticular}$), (G_{cystic} and G_{s_nodule}). For each such pair, we also calculate the uncertainty associated with the mean difference. It goes without saying that the larger the uncertainty in relation to the mean difference, the poorer the feature. These mean differences will then be utilized to set a weight vector for the feature.

Let's first focus on the pair (G_{cystic}, $G_{reticular}$). For pairwise comparisons, the Bonferroni approach can be used to construct uncertainty intervals for the individual feature components of the difference vector $\overline{X_{cystic}} - \overline{X_{reticular}}$. Let $N_t = N_{cystic} + N_{reticular}$ be the total number of sample vectors available. Under the condition that the confidence level is at least $(1-\alpha)$, we can obtain the following interval for the uncertainty in the difference of the mean values of the i^{th} feature:

$$(L_i, R_i) = \overline{X_{cystic,i}} - \overline{X_{reticular,i}} \pm$$
$$t_{N_t-2}(\alpha')\sqrt{\frac{w_{i,i}}{N_t-2}\left(\frac{1}{N_{cystic}} + \frac{1}{N_{reticular}}\right)} \qquad (7)$$

where $\alpha' = \frac{\alpha}{2p}$ and $w_{i,i}$ is the i^{th} diagonal element of $\mathbf{W}$ (defined in the previous section) and $t_{N_t-2}(\alpha')$ is the student t-distribution with $N_t - 2$ degrees of freedom. The size of this uncertainty interval is given by $R_i - L_i$. Evidently, when the second term in Eq. 7 is zero, there is no uncertainty in the difference of the mean values for feature i since L_i becomes equal to R_i. By the same token, when the second term in Eq. 7 is greater than the first, the uncertainty domi-

nates, making such a feature unreliable. The weight given to such a feature is zero. We only compute the weight for a feature if the second term of Eq. 7 is less than the first term for that feature.

The quality of the i^{th} feature for discriminating between the categories G_{cystic} and $G_{reticular}$ can now be measured by the following h factor:

$$h_{i,cystic,reticular} = \left| \frac{\overline{X_{cystic,i}} - \overline{X_{reticular,i}}}{\frac{R_i - L_i}{2}} \right| \quad (8)$$

These quality factors can be computed for the i^{th} feature for every pairing of G_{cystic} with the other categories. Subsequently, the quality factors can be combined into a single weight for the i^{th} feature:

$$w_{cystic,i} = \frac{\sum_{j \in \{reticular,s\_nodule\}} N_j h_{i,cystic,j}}{\sum_{k \in \{reticular,s\_nodule\}} N_k} \quad (9)$$

All such weights computed for the different feature components in this example are denoted by a vector of weights called W_{cystic} for this particular example. In general, for perceptual category g, this vector would be denoted W_g.

5 Experiments

We will present below some of our experiments with the two sets of features described previously. We used the general purpose features to validate the assumption that local features lead to significantly better retrieval results than global features. We also designed an experiment to compare the retrieval precision obtained when using the perceptual categories features against that when using the general features.

5.1 Local versus Global attributes

Table 1 shows for each disease category in our database the total number of queries for the category, 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 diagnoses as the query image. To assess the importance of local features versus global we used two different sets of attributes. The first is a combination of attributes extracted from the PBR region ($R_1(P)$) and attributes contrasting the PBR to the rest of the lung region (C). The second set of attributes ($R_2(G)$) was customized to a global approach to image characterization and were chosen by the SFS algorithm when optimizing performance for the entire lung region. We used the nearest-neighbor retrieval method after removing from the database the query-patient images.

5.2 Precision based on perceptual and disease categories

For this experiment, our database contained 610 PBRs from 314 HRCT lung images. We have two kinds of experimental results to report. The first, illustrated by Fig. 6, shows the retrieval precision with respect to just the perceptual categories. This experiment consists of the following steps: 1) Randomly select an image from the database as a query image; 2) Ask the system to retrieve four most similar images from the database taking into account the feature weights discussed in Section 4.7 for the different perceptual categories; and 3) Compare the perceptual category of the PBRs in the query image with the perceptual categories of the PBRs in the retrieved images. (Therefore, for these experiments we do not pay any attention to the disease labels associated with the PBRs.)

The retrieval precision taking into account the disease labels of the PBRs shown in Fig. 7. The steps that constitute this experiment are similar to those described above, except for the following three differences: 1) the retrieval precision is computed on the basis of the disease label of the query image vis-a-vis the disease labels of the retrieved images; 2) the image similarity metric is computed directly from the W_i weight vectors for i = 1,2,.., 9 for the nine perceptual categories (these vectors were defined at the end of Section 4.7); and 3) the image similarity metric takes into account the fact that in the database the distribution of the PBRs with respect to the perceptual categories is not uniform by associating the following weight with each perceptual category:

$$W_{i,updated} = \frac{[N_1 W_1,\ N_2 W_2, ..., N_{N_c} W_{N_c}]}{\sum_{i=1}^{N_c} N_i} \quad (10)$$

where N_i is the number of PBRs in the training data for perceptual category i.

On the average, using perceptual categories for retrieval in the manner described here resulted in improving the precision rates from 71.77% to 77.60% over the traditional method mentioned in section 3. Note that three out of the twelve disease categories experienced reduced precision with perceptual categories. We believe the problems are caused by the fact that for some of these diseases, such as panacinar (PA), the number of entries in the database is small compared to the entries for another disease, such as centrilobular emphysema (CLE).

6 Hierarchical Retrieval

One of the important problems in content-based image retrieval is the selection of the retrieval

Table 1: *Comparison of localized versus global attributes.*

Diagnosis	Query Images	Correct Retrievals $R_1(P)+C$	Correct Retrievals $R_2(G)$	Percent of Total $R_1(P)+C$	Percent of Total $R_2(G)$
CLE:	168	2.92 ± 0.18	2.12 ± 0.85	73	53
PSE:	29	3.04 ± 0.27	1.68 ± 1.07	76	42
BO:	28	3.00 ± 0.32	2.32 ± 0.55	75	58
HE:	18	2.92 ± 0.53	2.40 ± 0.33	73	60
MC:	12	2.64 ± 1.02	2.48 ± 1.21	66	62
PA:	16	2.80 ± 1.08	2.32 ± 1.32	70	58
PCP:	15	2.48 ± 0.85	2.40 ± 0.23	62	60
SA:	16	2.76 ± 0.71	1.96 ± 0.75	69	49
Total DB	302	2.89 ± 0.36	2.14 ± 0.82	72.3	53.6

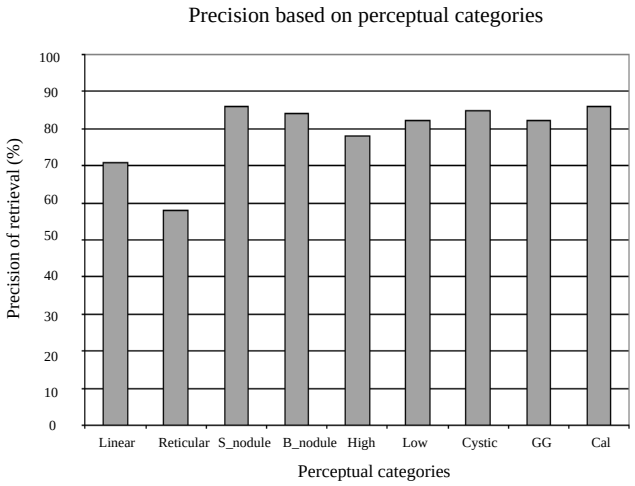


Figure 6: *Retrieval precision based on perceptual categories.*

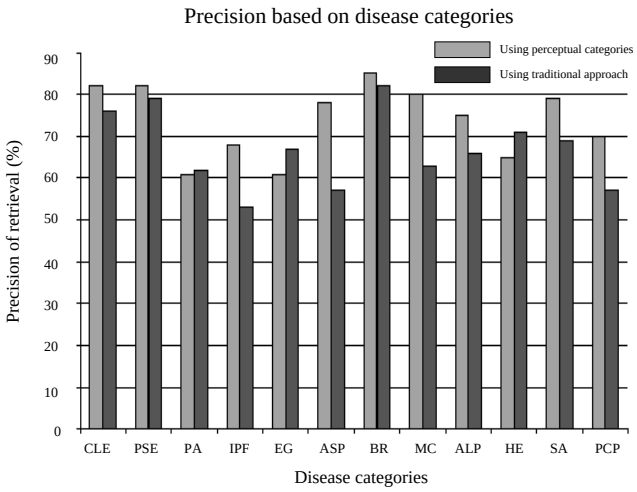


Figure 7: *Retrieval precision based on disease categories.*

method. In medical imaging it is often the case that each image belongs to a major class (disease) but images within each class can vary widely with respect to visual similarity (because of the severity of the disease). Thus, by first classifying the query image to a disease according to a set of features, and then retrieving the most similar images to the query image using the same set of features might not yield the most desirable results. This is because the features that differentiate among diseases might not be the most effective in retrieving visually similar images within a class.

For this purpose the Customized Queries Approach (CQA) was developed ([22, 1]). CQA works at two levels: first it finds the features that discriminate the major classes, and then it "customizes" the query by using the specialized set of features in the query's class to obtain the best n images. In order to find the set of features for the first level, the algorithm employs SFS ([13]). Finding the best set of features within each class is an unsupervised feature selection problem, and CQA resolves it by performing clustering and feature selection simultaneously.

More specifically, the basic idea is to search through feature subset space, evaluating each subset, F_t, by first clustering in space F_t using the EM [8] algorithm and then evaluating the resulting cluster using a chosen clustering criterion. The result of this search is the feature subset that optimizes the 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 search can be used [9], i.e. each time we add the feature that when combined with the current chosen set yields the largest improvement with respect to a separability criterion. For more details the reader is referred to [22] and [4].

Table 2 contains some comparative results between CQA and the traditional method, i.e. using the same features in all levels. SA stands for Strongly Agree, A stands for Agree, NS for Not Sure, D for Disagree and SD for Strongly Disagree.

Table 2: *Experimental results for Customized Queries.*

Disease	Traditional Method					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

7 Relevance Feedback

A retrieval system with relevance feedback prompts the user for feedback on retrieval results and then utilizes this feedback on subsequent retrievals in order to increase the retrieval performance. A popular approach in this area has been to use a weighted k nearest neighbors retrieval, where the weights are determined by a function of the user feedback. Systems that employ this strategy are MARS [16] and Probabilistic Feature Relevance Learning (PFRL) [19]. In MARS a feature's weight is determined by examining the feature's variance across the set of retrieved images marked as relevant by the user. In PFRL a feature's weight is computed by examining the k marked images closest to the query with respect to only that feature.

The relevance feedback mechanism of ASSERT, called Relevance Feedback Decision Trees (RFDT), was introduced in [23]. This work casts relevance feedback as a classification problem between the two classes *relevant* and *irrelevant* and relies on machine learning techniques to solve the problem efficiently. It operates as follows: on the first iteration, no feedback information exists, so the retriever performs an unweighted k nearest neighbors retrieval. The user then marks the k retrieved images as relevant or irrelevant as he or she sees fit. The query image is marked automatically relevant. This feedback is relayed back to the system and the second iteration begins. Now the $k+1$ images are viewed as training data belonging to the classes relevant and irrelevant and a decision tree is induced using C4.5 [20]. Once the tree is formed, it is used to select the next set of k images to present to the user. To this end, the entire database of feature vectors is classified via the learned tree, and the images classified as *relevant* are assembled in a list. From this list, the k images closest to the query image are returned using k nearest neighbors retrieval. The graphs in Fig. 7 show the performance of RFDT against PFRL. For more information the reader is referred to [23].

8 Summary

Content based image retrieval has become an important area in computer vision. Much has been accomplished, but much more remains to be done. In this paper, we have highlighted some of the problems unique to automated retrieval from large medical image databases and presented solutions to some of them in the specific context of HRCT images of the lung.

References

[1] C. E. Brodley A. Kak J. Dy C. Shyu A. Aisen and L. Broderick. Content-based retrieval from medical image databases: A synergy of human interaction, machine learning and computer vision. In *Proc. of the 16th National Conference on Artificial Intelligence*, 1999.

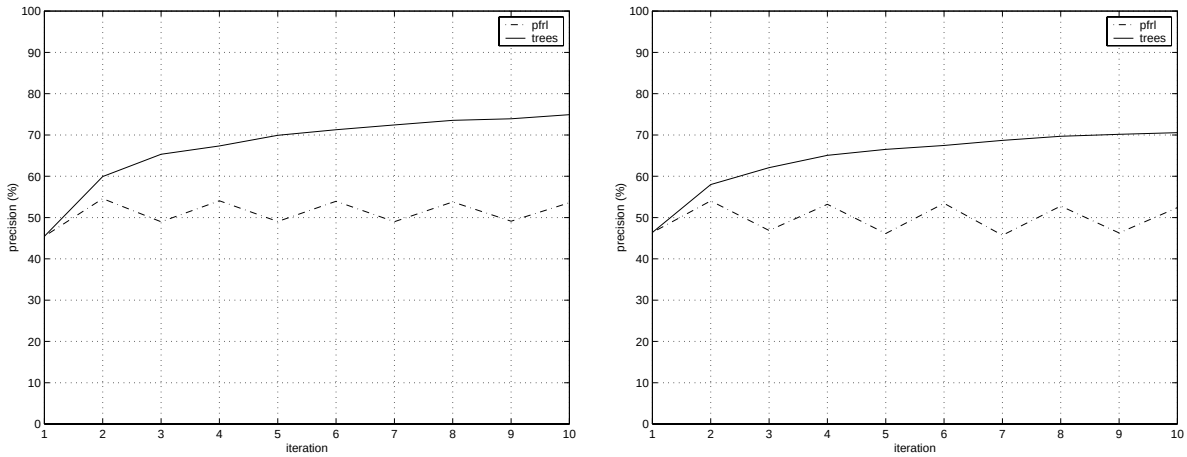
[2] C. Shyu C. Brodley A. Kak A. Kosaka A. Aisen and L. Broderick. ASSERT: A physician-in-the-loop content-based image retrieval system for HRCT image databases. *Computer Vision and Image Understanding*, 75(1/2):111–132, 1999.

[3] C. Shyu A. Kak C. Brodley and L. Broderick. Testing for human perceptual categories in a physician-in-the-loop CBIR system for medical imagery. In *Workshop of Content-Based Access of Image and Video Databases*, pages 18–22. IEEE, 1999.

[4] J. Dy C. Brodley. Feature subset selection and order identification for unsupervised learning. In *Proc. of the 17th International Conference on Machine Learning*, 2000.

[5] P.M. Kelly T.M. Cannon and D.R. Hush. Query by image example: The CANDID approach. In *Storage and Retrieval for Image and Video Databases III*, pages 238–248. SPIE Vol. 2420, 1995.

[6] W. W. Chu C. C. Hsu A. F. Cardenas and R. K. Taira. A knowledge-based image retrieval with spatial and temporal constructs. *IEEE Trans-*



(a) $k = 4$ (b) $k = 10$

Figure 8: *Average retrieval precision of PFRL (dotted line) and RFDT(solid line) for our HRCT lung database.*

actions on Knowledge and Data Engineering, 10(6):872–888, 1998.

[7] S. Sclaroff L. Taycher M La Cascia. Imagerover: A content-based image browser for the World Wide Web. In *Workshop on Content-based Access of Image and Video Libraries*. IEEE, 1997.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Ser. B*, 39(1):1–38, 1977.

[9] K. Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition*. Academic Press, Inc., 1990.

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

[11] R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1998.

[12] J. Kittler. Feature set search algorithms. In *Pattern Recognition and Signal Processing*, pages 41–60, 1978.

[13] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[14] M. Ortega Y. Rui K Chakrabarti S. Mehrotra and T. Huang. Supporting similarity queries in MARS. In *5th Int. Multimedia Conference*. ACM, 1997.

[15] W. R. Webb N. L. Muller and D. P. Naidich. *High-Resolution CT of The Lung*. Lippincott-Raven, Philadelphia, 1996.

[16] Y. Rui T. Huang M. Ortega and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. In *Proc. of the Int. Conference on Image Processing*. IEEE, 1997.

[17] A. Pentland R. W. Picard and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. In *Storage and Retrieval for Image and Video Databases*, pages 34–47. SPIE, 1994.

[18] F. Korn N. Sidiropoulos C. Faloutsos E. Siegel Z. Protopapas. Fast and effective retrieval of medical tumor shapes. *IEEE Transactions on Knowledge and Data Engineering*, 10(6):889–904, 1998.

[19] J. Peng B. Bhanu S. Qing. Probabilistic feature relevance learning for content-based image retrieval. *Computer Vision and Image Understanding*, 75:150–164, 1999.

[20] J. R. Quinlan. *Programs for Machine Learning*. Morgan Kaufman, San Mateo, CA, 1993.

[21] A. Rosenfeld and A. C. Kak. *Digital Picture Processing*. Academic Press, 1982.

[22] J. Dy C. Brodley A. Kak C. Shyu and L. Broderick. The customized-queries approach to CBIR using EM. In *Computer Vision and Pattern Recognition*, pages (II)400–406. IEEE, 1999.

[23] S. D. MacArthur C. E. Brodley C. Shyu. Relevance feedback decision trees in content-based image retrieval. In *Workshop on Content-Based Access of Image and Video Libraries*. IEEE, 2000.

[24] M. Flickner H. Sawhney W. Niblack J. Ashley Q. Huang B. Dom M. Gorkani J. Hafner D. Lee D. Petkovic D. Steele P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, pages 23–32, 1995.