This paper appears in: IEEE Workshop on Content-Based Access of Image and Video Libraries, 1999

Testing for Human Perceptual Categories in a Physician-in-the-loop CBIR System for Medical Imagery

Chi-Ren Shyu, Avi Kak, Carla E. Brodley
{chiren, kak, brodley}@ecn.purdue.edu
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907

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

Abstract

We have addressed the following question in this contribution: To what extent should the domain experts, in our case physicians, be believed with regard to what they claim to see in images that allows them to recognize different types of pathology?

Until recently our approach was to have a physician delineate the pathology bearing regions in the images. We then used what could be referred to as a scattershot approach to the characterization of these regions, meaning that we'd extract a very large number of features from these regions. Subsequently, we'd reduce the dimensionality of this feature space by using standard search techniques, such as the Sequential Forward Selection method.

This contribution represents an alternative to the scattershot approach to initial feature extraction. In this paper, we first describe the perceptual categories that the physicians claim to use for classifying images as belonging to different diseases. We then describe the specific low-level features that need to be extracted to determine the presence or the absence of the various perceptual categories. We subsequently show the discriminatory power of the perceptual categories by presenting retrieval results obtained when a query image is matched with the database images on the basis of the presence or the absence of the various perceptual categories.

Keywords: Medical image databases, CBIR, feature testing, human perception.

1 Introduction

In the past few years, researchers have been working on finding the right image features for the domain of CBIR systems in medicine [2, 4, 6, 8, 11, 12]. While some of these systems [2, 6, 8] use specialized features, such as the symmetry properties of the ventricular line in the MR images of the brain, others, including ours [11, 12], employ what may be referred to as the scattershot approach in which an exhaustive set of low-level features is extracted for the characterization of image pathology. The dimensionality of the feature space is subsequently reduced by searching for a representative subset using greedy algorithms such as the Sequential Forward Selection search [5], so that eventually one retains only those features that are maximally discriminatory with regard to the different diseases.

The work reported in this paper is motivated by our desire to test directly the information content and the discriminatory power of the perceptual categories that the physicians claim they are using for diagnosis and retrieval in the domain of high-resolution computed tomography (HRCT) images of the lung. Rather than initially extract an exhaustive set of low-level features and then reduce the dimensionality of this space, we now want to extract principally only those low-level features that measure the presence or the absence of the perceptual categories used by the physicians.

In this paper, in Section 2 we first delineate the perceptual categories used by physicians for recognizing lung pathology in the HRCT domain. We also include in this discussion a list of the low-level features that are appropriate for measuring the presence or the absence of each perceptual category. In Section 4, we discuss how we use the hypothesis testing aspect of MANOVA to find the subsets of low-level features that are best able to discriminate among the different perceptual categories. In Section 6, we show how the low-level features should be weighted to increase the measured "separation" among the various perceptual categories. Finally, in Section 7 we present retrieval results using the physicians' perceptual categories and compare them to the results using our earlier scattershot approach.

2 Physicians' Perceptual Categories for Recognizing Lung Pathology

Fig. 1 shows the perceptual categories that physicians use for describing the visual structure of a pathology bearing region (PBR) in an HRCT image. The four major categories are [13]: *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. 1 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

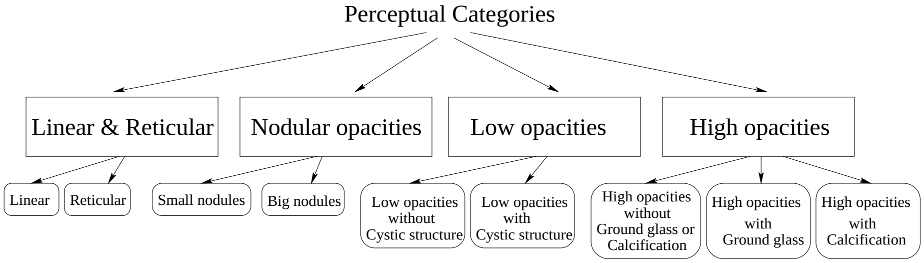

Perceptual Categories
Linear & Reticular
Nodular opacities
Low opacities
High opacities
Linear
Reticular
Small nodules
Big nodules
Low opacities without Cystic structure
Low opacities with Cystic structure
High opacities without Ground glass or Calcification
High opacities with Ground glass
High opacities with Calcification

Figure 1: *Perceptual categories used by physicians.*

visual structure inside the PBR would either be linear or reticular, corresponding to the two leaf nodes in Fig. 1. The rest of this section provides further details regarding the visual structure associated with these categories.

2.1 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. 2(a)) and peripheral honeycombing (Fig. 2(b)). Since the walls of the bronchi are characterized by adjacent low and high attenuation regions, they can be extracted by dual-thresholding [9]. 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.*

2.2 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. 2(d). For the case of small nodules, their distribution carries diagnostic information. When the distribution is random, then the nodules appear widely and evenly throughout the lung as shown in Fig. 2(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* [10], and the *gray-level mean of nodules.*

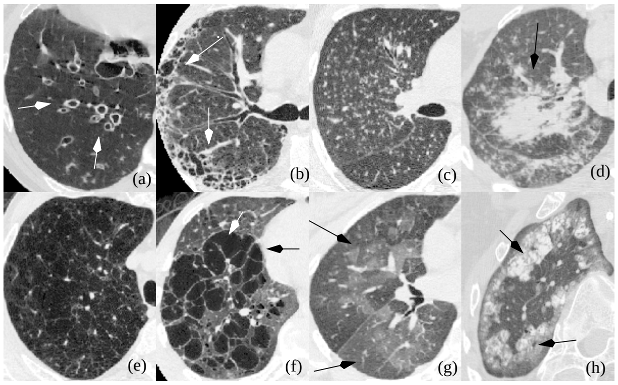


Figure 2: *Visual information for different lung pathologies: (a) Bronchial structure (linear & reticular), (b) Honeycombing (linear & reticular), (c) small nodules (nodular opacities), (d) big nodules (nodular opacities), (e) low-attenuation (low opacities), (f) cystic structure (low opacities), (g) ground-glass (high opacities), (h) calcification (high opacities).*

2.3 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. 2(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. 2(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 [1] 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*.

2.4 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. 2(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. 2(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.

The discussion so far has identified a set of gray-level thresholds that are used to extract the attributes corresponding to the relevant perceptual categories. How to set these thresholds is obviously an important issue in the design of a CBIR system. Each threshold is chosen by applying Otsu's algorithm [7] to the relevant histograms. This algorithm is based on the assumption that a histogram is a mixture of two Gaussian classes and that the optimum threshold that separates them is the ratio of between class variance and the sum of within class variances. This approach allows each threshold to adapt to the image in question.

3 The Special Database Constructed For Ascertaining The Significance Of The Perceptual Categories

We asked one of the physicians participating in our research program to mark our HRCT database images with regard to the presence or the absence of the various perceptual categories. A special graphical interface tool was devised for this purpose. Using this tool, for each database image the physician could check as many of perceptual categories as applicable to the PBRs in that image. The distribution of the images in the resulting database will be shown in the section on experimental results.

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

We have used multivariate analysis of variance (MANOVA) [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. 1. 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 $N_T - N_g (= N_{\mathrm{rest}})$ 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_{rest} samples will be denoted X_{rest} and the mean of this data by $\overline{X_{\mathrm{rest}}}$

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^{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^{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:

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

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

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

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

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 $\mathbf{H_0} : \tau_{\mathrm{g}} \;=\; \tau_{\mathrm{rest}}$, 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. τ_{rest} denotes $\overline{X_{\mathrm{rest}}} - \overline{X}$. To test the $\mathbf{H_0}$ hypothesis, we first compute *Wilks' lambda* Λ^* [14]:

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

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 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

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

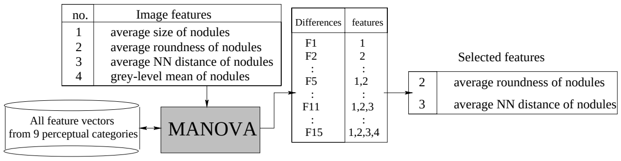
to reject hypothesis H_0 at confidence level $(1 - \alpha)$. $F_{d_B, d_W}(\alpha)$ is the upper $100\alpha\%$ of the F-distribution with d_B and d_W degrees of freedom.

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.

5 Choosing the Maximally Discriminatory Feature Set

In Section 2, 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.[1] 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. 3. 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 was used to determine the quality of each feature combination. We selected the feature combination that corresponds

[1] We certainly admit the possibility that there might be other low-level features not yet imagined by us that could prove to be powerful detectors of the perceptual categories. To test the efficacy of the low-level features of Section 2, we have performed experiments by also including generic features commonly used for measuring texture, gray scale, edginess, etc. Although these experiments have reinforced our belief in the appropriateness of our low-level features, we remain open to the possibility that there might exist other low-level features that would be superior detectors of the perceptual categories. Space limitations prevent us from discussing here the details of these experiments, save for mentioning that the generic features we used in such studies are discussed in greater detail in [12].



no.	Image features
1	average size of nodules
2	average roundness of nodules
3	average NN distance of nodules
4	grey-level mean of nodules

Differences	features
F1	1
F2	2
:	:
F5	1,2
:	:
F11	1,2,3
:	:
F15	1,2,3,4

Selected features

2	average roundness of nodules
3	average NN distance of nodules

Figure 3: *Choosing the maximally discriminatory features for the small nodule perceptual category.*

to the highest F value. Fig. 3 shows this process pictorially. In this case, the subset {2,3} produced the highest F value. This process is repeated for each perceptual category.

6 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 dominates, 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:

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

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:

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

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.

7 Experimental Results

Our database, created in the manner described in Section 3, contains 610 PBRs from 314 HRCT lung images. Table 1 shows the database distribution with respect to the different diseases and with respect to the different perceptual categories. The diseases in the database are centrilobular emphysema (CLE), paraseptal emphysema (PSE), panacinar (PA), idiopathic pulmonary fibrosis (IPF), eosinophilic granuloma (EG), aspergillus (ASP), bronchiectasis (BR), metastatic calcification (MC), alveolar proteinosis (ALP), hemorrhage (HE), sarcoid (SA), and pneumocystis carinii pneumonia (PCP).

7.1 Precision based on perceptual and disease categories

We have two kinds of experimental results to report. The first, illustrated by Fig. 4, 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

Table 1: *Distributions of lung diseases and perceptual categories in our database.*

Disease	PBR/Image	Linear	Reticular	S_Nodule	B_Nodule	High	Low	Cystic	GG	Cal
CLE	326/158	0	0	0	0	0	326	0	0	0
PSE	54/28	0	0	0	0	0	54	0	0	0
PA	10/6	0	0	0	0	0	10	0	0	0
IPF	51/24	15	34	0	0	30	20	21	0	0
EG	58/29	5	5	0	0	0	58	57	0	0
ASP	16/16	0	0	0	16	16	0	0	0	0
BR	28/14	0	0	0	0	12	14	12	12	0
MC	8/5	0	0	1	7	8	0	0	8	8
ALP	15/8	4	2	1	0	14	0	0	3	0
HE	10/5	0	10	0	0	10	0	0	2	0
SA	15/11	0	0	15	6	8	0	0	0	0
PCP	19/10	0	0	0	0	19	0	0	5	0
Total	610/314	24	51	17	29	117	482	90	30	8

most similar images from the database taking into account the feature weights discussed in Section 6 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 for each perceptual category is shown in Fig. 4. The precision for the cases of *S_nodule*, *B_nodule*, *Cystic*, and *Ground_glass* is reasonably good. The relatively poor performance for the *Linear* and *Reticular* categories can be attributed to the lack of purity of these categories, meaning that the PBRs that exhibit these categories also exhibit other categories simultaneously.

The retrieval precision taking into account the disease labels of the PBRs shown in Fig. 5. 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 6); 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:

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

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

The precision of the retrieval results is shown in Fig. 5. 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. The dark bars in this figure correspond to using the "traditional" approach described in [12] in which we start with an exhaustive list of low-level image features that are subsequently pruned by employing the sequential forward selection (SFS) method [5]. 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 small number of entries in the database are overshadowed completely by the entries for another disease, such as centrilobular emphysema (CLE). We are still investigating the matter of this non-uniformity of results across the various diseases and as to what extent this is a function of the non-uniform nature of the database.

8 Conclusion

What specific features to use for content-based retrieval is more a function of the level of ingenuity of researchers than a result of some precise scientific analysis. In the past, we used the scattershot approach that consisting of characterizing the pathology bearing regions with an exhaustive set of features and then using a standard dimensionality reduction tool to pull out the feature set that appeared to be maximally discriminatory with respect to the disease categories. In this paper, we have eschewed this previous approach. Instead, we have tried to extract only those low-level features that had the potential of measuring the presence or the absence of the various perceptual categories that the physicians claim to use for disease diagnosis. MANOVA was then used to select a subset of these features that was maximally discriminatory with respect to the perceptual categories. According to our preliminary experimental results, this new approach to feature extraction and image characterization has yielded retrieval performance that appears to be superior for some disease classes and inferior for other disease classes. As to why that is the case is a topic of ongoing investigation.

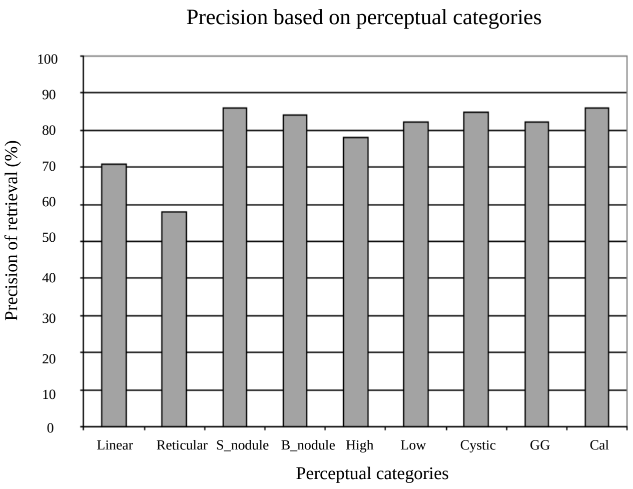


Figure 4: *Retrieval precision based on perceptual categories.*

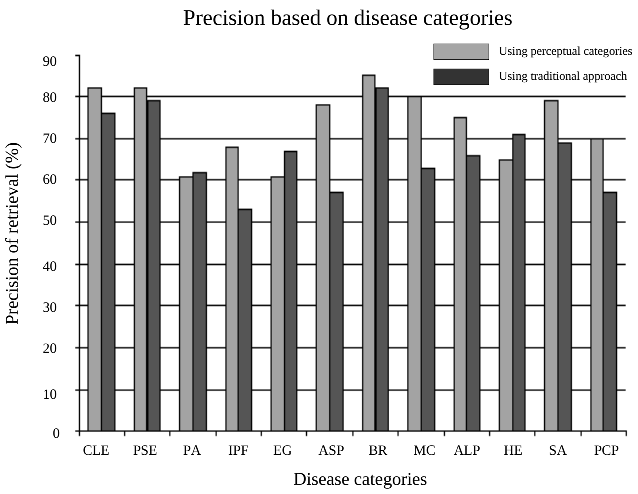


Figure 5: *Retrieval precision based on disease categories.*

9 Acknowledgments

This work is supported by the National Science Foundation under Grant No. IRI9711535 and the National Institute of Health under grant No. 1 R01 LM06543-01A1. We also wish to acknowledge fruitful discussions with Professor George McCabe and James O'Malley of the Department of Statistics at Purdue University.

References

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

[2] C. C. Hsu, W. W. Chu, and R. K. Taira, A knowledge-based approach for retrieving images by content, *IEEE Trans. on Knowledge and Data Engineering*, Vol. 8, No. 4, pp. 522-532, 1996.

[3] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, fourth Edition, Prentice Hall, 1998.

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

[5] J. Kittler, Feature set search algorithms. In: C. H. Chen, Ed. , *Pattern Recognition and Signal Processing* pp. 41 - 60. Sijthoff and Noordhoff, Alphen ann den Rijn, The Netherlands, 1978

[6] 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*, Vol. 10, No. 6, pp. 889-904, 1998.

[7] N. Otsu, A threshold selection method from gray-level histogram, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-9, No. 1, pp.62-66, 1979.

[8] G. P. Robinson, H. D. Tagare, J. S. Duncan, and C. C. Jaffe, Medical image collection indexing: shape-based retrieval using KD-tree, *Computerized Medical Imaging and Graphics*, Vol. 20, No. 4, pp. 209-217, 1996.

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

[10] H. Schwarz and H. E. Exner, The characterization of the arrangement of feature centroids in plans and volumes. *J. Microsc,* pp. 129-155, 1983.

[11] C. R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. Aisen and L. Broderick, Local versus global features for content-based image retrieval, *Proc. IEEE Workshop of Content-Based Access of Image and Video Databases*, Santa Barbara, CA, June 1998.

[12] C. R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen and L. S. Broderick, ASSERT: A physician-in-the-loop content-based image retrieval system for HRCT image databases, to appear in *Computer Vision and Image Understanding, Special Issue on Content-Based Access of Image and Video Libraries*, 1999.

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

[14] Wilks, S. S., Certain generalizations in the analysis of variance, *Biometrika,* 24, pp. 471-494, 1932.