Content-Based Image Retrieval for Pulmonary Computed Tomography Nodule Images

Michael Lam[a], Tim Disney[b], Mailan Pham[c], Daniela Raicu[d], Jacob Furst[d], Ruchaneewan Susomboon[d]
[a]James Madison University, Harrisonburg, VA, USA, 22807
[b]Seattle Pacific University, Seattle, WA, USA, 98119
[c]Mt. Holyoke College, South Hadley, MA, USA, 01075
[d]Intelligent Multimedia Processing Laboratory
School of Computer Science, Telecommunications, and Information Systems
DePaul University, Chicago, IL, USA, 60604

This material is based upon work supported by the National Science Foundation under Grant No. 0453456.

ABSTRACT

Research studies have shown that advances in computed tomography (CT) technology allow better detection of pulmonary nodules by generating higher-resolution images. However, the new technology also generates many more individual transversal reconstructions, which as a result may affect the efficiency and accuracy of the radiologists interpreting these images.

The goal of our research study is to build a content-based image retrieval (CBIR) system for pulmonary CT nodules. Currently, texture is used to quantify the image content, but any other image feature could be incorporated into the proposed system. Unfortunately, there is no texture model or similarity measure known to work best for encoding nodule texture properties or retrieving most similar nodules. Therefore, we investigated and evaluated several texture models and similarity measures with respect to nodule size, number of retrieved nodules, and radiologist agreement on the nodules' texture characteristic.

The results were generated on 90 thoracic CT scans collected by the Lung Image Database Consortium (LIDC). Every case was annotated by up to four radiologists marking the contour of nodules and assigning nine characteristics (including texture) to each identified nodule. We found that Gabor texture descriptors produce the best retrieval results regardless of the nodule size, number of retrieved items or similarity metric. Furthermore, when analyzing the radiologists' agreement on the texture characteristic, we found that when just two radiologists agreed, the average precision increased from 88% to 96% for both Gabor and Markov texture features. Moreover, once three or four radiologists agreed the precision increased to nearly 100%.

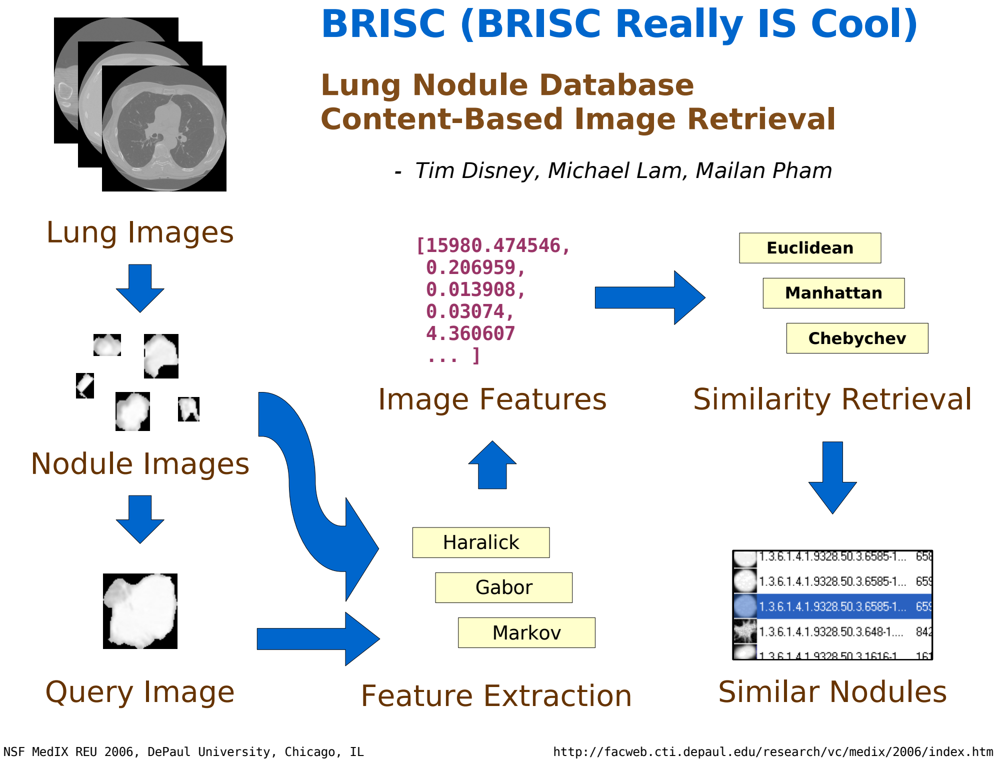
Keywords: Content-based image retrieval, texture feature, co-occurrence matrix, Gabor filter, Markov random field

Introduction

Lung cancer causes more deaths each year than the three next most common cancers (colon, breast and prostate) combined, and it is estimated that there were over 160,000 deaths in the United States due to lung cancer in 2006.[1] Lung cancer should be treated as early as possible, but it is hard to detect using conventional radiography. Computed tomography (CT) scanning has been found to increase the detection rate of pulmonary nodules.[2] However, there is still much to improve in computer-assisted diagnosis (CAD) systems, particularly in the area of nodule comparison and retrieval.

In this paper, we present a content-based image retrieval (CBIR) system for pulmonary nodule lookup. We also examine and compare several different texture-based image comparison methods.

Techniques for texture analysis are normally grouped under one of four categories: structural, statistical, transform, and model-based. Structural approaches seek to understand the hierarchal structure of the image, while statistical methods describe the image using pure numerical analysis of pixel intensity values. Transform



BRISC (BRISC Really IS Cool)

Lung Nodule Database
Content-Based Image Retrieval

- Tim Disney, Michael Lam, Mailan Pham

Lung Images

Nodule Images

Query Image

[15980.474546,
0.206959,
0.013908,
0.03074,
4.360607
...]

Image Features

Haralick

Gabor

Markov

Feature Extraction

Euclidean

Manhattan

Chebychev

Similarity Retrieval

1.3.6.1.4.1.9328.50.3.6585-1... 658
1.3.6.1.4.1.9328.50.3.6585-1... 659
1.3.6.1.4.1.9328.50.3.6585-1... 659
1.3.6.1.4.1.9328.50.3.648-1.... 842
1.3.6.1.4.1.9328.50.3.1616-1 161

Similar Nodules

NSF MedIX REU 2006, DePaul University, Chicago, IL

http://facweb.cti.depaul.edu/research/vc/medix/2006/index.htm

Figure 1. System Overview

approaches generally perform some kind of modification to the image, obtaining a new "response" image that is then analyzed as a representative proxy for the original image, and model-based methods are based on the concept of predicting pixel values based on a mathematical model.

Although the structural approach represents the image's texture well by demonstrating the hierarchal structure of the texture, this feature is more useful for texture synthesis than for texture analysis.[3] Therefore, we are only focusing on texture feature extraction methods from the three latter categories: co-occurrence matrices (statistical), Gabor filters (transform) and Markov Random Fields (model-based).

These methods are used to extract a feature vector that represents an image's signature. This vector is then compared with the vectors of other images using various similarity measures. We have created a basic prototype CBIR system for querying lung nodules using the three texture models (co-occurrence, Gabor and Markov) and five similarity measures (Euclidean, Chebyshev and Manhattan for co-occurrence, as well as Chi-Squared and Jeffrey Divergence for Gabor and Markov). The diagram of the system is shown in Figure 1.

Related Work

The first known large-scale comparison of texture features was done by Ohanian and Dubest in 1992. They tested 16 co-occurrence features, 4 Markov Random Field (MRF) features, 16 Gabor filter features, and 4 fractal geometry features on 3200 32x32 sub-images and found that co-occurence performed the best.[4] However, while Ohanian and Dubest evaluated the feature types in respect to their ability to classify texture correctly, we seek to evaluate the features by in their performance in image retrieval. Deselaers et al. compared texture features for two different image retrieval tasks (color photographs and medical radiographs): pixel-value, color histograms, invariant feature histograms, Gabor feature histograms, Tamura texture feature histograms, local features direct transfer, and region based features. They found that for the databases of medical radiographs, using the pixel values directly as features results in the best retrieval performance.[5] Thus far, there have been no papers published in which Haralick co-occurence, MRF, and Gabor filters are evaluated in their performance in a CBIR system for medical images.

There are several other CBIR projects currently underway in the medical field[6] and particularly with lung CT images. The largest, ASSERT, is being developed at Purdue University and was first published in 1999. It proposed a "physician-in-the-loop" system in which a radiologist highlighted pathology-bearing regions and then the system ran a query for images with similar regions. The system used a variety of different image features, including co-occurrence statistics, shape descriptors, Fourier transforms and global gray level statistics. The system also utilized physician-provided ratings of features such as homogeneity, calcification and artery size. Two methods were tested for similarity comparison: nearest-neighbor and multidimensional hashing. The latter proved faster although a bit less precise. The best precision reported by the system was 76.3%.[7–9]

There was also a lung CT CBIR system developed at Taichung Veteran's General Hospital in Taiwan (published in 2001), which segmented the image into blocks and used a Kohonen neural network to classify the blocks and return relevant images, obtaining an error rate of 0.14%.[10] A more recent project was published in 2004 at the University of Tokushima in Japan, which used shape descriptors and density histograms to classify and retrieve 3D lung CT volumes. Only preliminary work has been published so there are no precision or recall metrics for this system yet.[3]

Aside from the systems described above, there have been many image classification projects using CT lung images, such as a project at the Royal Brompton Hospital in London that used co-occurrence descriptors along with statistical moment features and acquisition-length parameters. These descriptors were analyzed with a supervised Bayesian classifier to classify various images of lung tissue as containing various pathologies. This system achieved a sensitivity of 73.6% and a specificity of 91.2%.[11]

There have also been advances made in the areas of segmentation, automated nodule detection and computer-aided diagnosis (CAD), such as a project at Chungnam National University in South Korea. This project experimented with different algorithms for lung segmentation and achieved a 96% sensitivity with no false positives.[12] Another project at the University of Occupational and Environmental Health School of Medicine in Japan used an artificial neural network to analyze physician-extracted clinical parameters and classify pulmonary nodules as either benign or malignant. The project used receiver-operating characteristic (ROC) curves to analyze the resulting true- and false-positive fractions. The best area index (A_z) value obtained was .951.[13] More recently, a project in Iran experimented with various methods of region-of-interest (ROI) extraction and achieved a best average classification rate of 91%.[14]

However, there are still many problems associated with content-based retrieval of medical images, such as the open nature of segmentation research and the large variability of feature selection as well as the lack of standardized toolkits and evaluation methods for medical CBIR systems.[6, 15, 16]

Texture Feature Extraction

Co-occurrence Matrices

Statistical methods such as Haralick co-occurrence matrices generally focus on the distributions and relationships of the gray levels in an image.[17]

The general idea of a co-occurrence matrix is to represent an image's texture features by counting pixel intensity pairs, using a matrix to keep track of all the pixel-pair counts. Our method calculates a separate co-occurrence matrix for each direction (0°, 45°, 90° and 135°) and displacement (1, 2, 3 and 4 pixels).

Here is an example matrix:

$$\begin{array}{ccccc}
0 & 0 & 1 & 2 & 1 \\
0 & 2 & 2 & 0 & 2 \\
1 & 1 & 1 & 2 & 1 \\
0 & 2 & 0 & 1 & 0 \\
0 & 1 & 2 & 2 & 0
\end{array}$$

This is the corresponding co-occurrence matrix, taken at a 0° angle and a one-pixel displacement:

$$\begin{matrix} 1 & 3 & 3 \\ 1 & 2 & 3 \\ 3 & 2 & 2 \end{matrix}$$

An finally, this is yet another co-occurrence matrix of the same image, taken at a 0° angle but with a two-pixel displacement:

$$\begin{matrix} 2 & 1 & 3 \\ 0 & 3 & 2 \\ 2 & 1 & 1 \end{matrix}$$

After the co-occurrence matrices are formed, Haralick features are calculated from the matrix data.[18] Since there are four directions, four displacements and eleven features, the result is a 4x4x11 matrix, which is averaged by distance. The minimum values by direction are then stored as eleven elements in the feature space. These elements can then be combined to form feature vectors of varying lengths. Since there are eleven features, there are $\sum_{k=1}^{11} \frac{11!}{k!(11-k!)} = 2047$ unique vectors (combinations of features).

To determine which of these vectors was best for our data set, we wrote a routine to perform a simulated query for each image in the database and calculate the mean precision and recall. We could then run this routine with various feature vectors and similarity measures to determine the best combination of query parameters for our set. We wanted to try all 2047 combinations with three different similarity measures (Euclidean, Manhattan and Chebychev) and five different numbers of retrieved images (1, 2, 3, 5 and 10 retrieved images).

Unfortunately, this would have required us to run our routine $2047(3)(5) = 30,705$ times, which would have taken a prohibitively long time. So we ran all 2047 combinations with only one similarity measure (Euclidean) and one number of retrieved image (five), to get a general idea of how well different feature vectors performed. After the initial trials, we chose the best 200 vectors to run with all similarity measures and numbers of returned images. This reduced the total number of trials to $200(3)(5) = 3000$ instead of 30,705.

Gabor Filters

In contrast to the statistical based co-occurrence matrix method, Gabor filtering is a transform based method of extracting texture information. The use of Gabor filters is motivated by Gabor filtering being "strongly correlated with the human visual system."[19] Gabor filters have also been successfully used in a number of other projects to extract texture information in order to perform similarity retrieval,[19,20] as well as texture segmentation.[21,22]

Gabor filtering is a way of extracting feature information from an image in the form of a response image. Several filters with varying parameters are applied to an image to acquire the response. A Gabor filter is a sinusoid function modulated by a Gaussian. The filters we used are defined by the following equation:

$$G(x,y) = e^{\left(\frac{-x_\theta^2-\gamma^2 y_\theta^2}{\sigma^2}+\frac{2\pi x_\theta i}{\lambda}\right)} \tag{1}$$

where

$$x_\theta = x\cos(\theta) + y\sin(\theta) \tag{2}$$

$$y_\theta = -x\sin(\theta) + y\sin(\theta) \tag{3}$$

and σ is the standard deviation of the Gaussian function, λ is the wavelength of the harmonic function, θ is the orientation, and γ is the spatial aspect ratio which is left constant at $\frac{1}{2}$. The spatial frequency bandwidth is the ratio σ/λ and is held constant and equal to .56. Thus there are two parameters which change when forming a

Gabor filter - θ and λ. The form of this equation and all constants are similar to the work done by T. Andrysiak et al.[20]

The size of our Gabor filters was set constant at 9x9 for simplicity. Once we have a Gabor filter, it is convolved with the original image to create a Gabor image response. Based on the work done by Andrysiak et al,[20] we are using only the odd component of the Gabor filter which does not produce imaginary output:

$$\psi_o(x, y) = exp\left(\frac{-x_\theta^2 - \gamma^2 y_\theta^2}{\sigma^2}\right) \sin\left(\frac{2\pi x_\theta}{\lambda}\right) \tag{4}$$

We convolve the image with 12 Gabor filters tuned to four orientations (θ) and three frequencies ($1/\lambda$). Figure 2 visualizes what happens to the Gabor filter when the orientation parameter is changed. Orientation varied from 0 to $3\pi/4$ (stepping by $\pi/4$) and frequency varied from .3 to .5 (stepping by .1).

Markov Random Fields

Markov Random Fields (MRFs) capture the local contextual information of an image.[23] The application of MRFs to extract textual information was first done by Jain and Cross in 1983.[17] Since then, MRFs have gained increasing popularity because of their ability to create an image model that can be successfully used for image classification, segmentation, and texture synthesis.[24]

In a MRF model, the image is represented by a two-dimensional lattice. The value at each pixel in the lattice is a random variable. For gray scale images, with 256 gray levels, each random variable can take on a value in the set {0, 1, 2, ..., 255}.[24] The lattice S with neighborhood system δ_s is said to be a MRF if for all $s \in S, p(X_s|X_r \text{ for } r \neq s) = p(X_s|X_{\delta_r})$, where X is a random variable.[25]

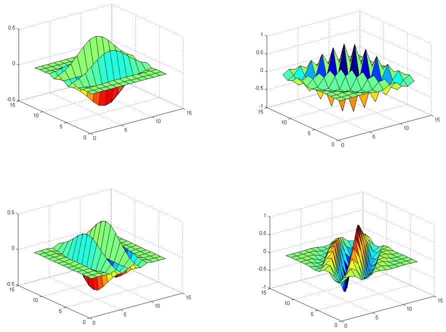


Figure 2. Visualization of Gabor filters varying by θ parameter. Top left $\theta = 0$, top right $\theta = \pi/4$, bottom left $\theta = \pi/2$, bottom right $\theta = 3\pi/4$

In order for a site to be a MRF, it must exhibit *Markovianity*, which describes the situation in which the value of each random variable is dependent only on its neighbors.[23] For instance, if the random variable X represents intensity values, the gray level of a pixel at X must depend on the gray levels of its neighbors.[26]

In a Gaussian Markov Random Field (GMRF), the image is represented on a local conditional probability distribution that is assumed to be Gaussian.[27] The four parameters for a GMRF model correspond to the four orientations between a neighboring pixel pair.[4] To extract five feature vectors for our CBIR system, we used an algorithm devised by Cesmeli: first estimate the four GMRF parameters, then derive four new features (as well as variance) from the estimated parameters.[27]

We used least-square estimation to estimate a set of four parameters for a second order GMRF model: $\hat{\Theta} = [\hat{\theta}_1\ \hat{\theta}_2\ \hat{\theta}_3\ \hat{\theta}_4]^T$, where $\hat{\theta}_1$ corresponds to 0° direction, $\hat{\theta}_2$ corresponds to the the 90° direction, $\hat{\theta}_3$ corresponds to the 45° diagonal, and $\hat{\theta}_4$ corresponds to the 135° diagonal.

$$\hat{\Theta} = \left[\sum_{r, r\pm\tau_j \in R(s)} Q(r)Q(r)^T\right]^{-1} \left[\sum_{r, r\pm\tau_j \in R(s)} Q(r)y_r\right] \tag{5}$$

Where $Q(r) = [(y_{r+\tau_1} + y_{r-\tau_1}), ..., (y_{r+\tau_4} + y_{r-\tau_4})]^T$, τ stands for the orientation, 0°, 90°, 45°, and 135°, respectively, r is the pixel location in the image, and $R(s)$ is the estimation window.

For example, the first scalar of the four scalars in $Q(r)$ corresponds to the 0° direction. In this case, $y_{r+\tau_1} + y_{r-\tau_1}$ is the sum of the two intensity values of the neighbor pixels that are to the left and the right of the pixel at location r.

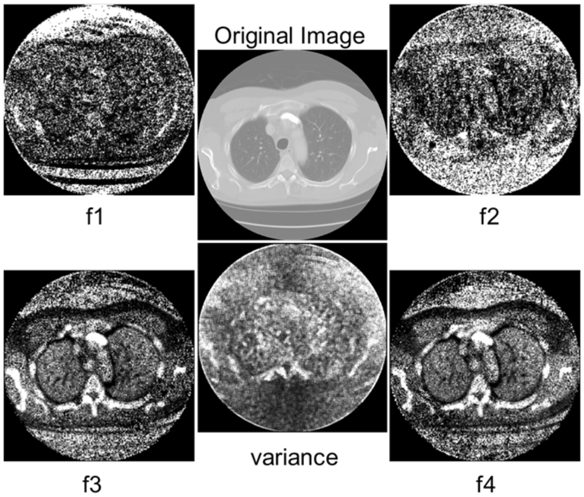


Figure 3. Visualization of a CT lung image and its five feature vector images: f1, f2, f3, f4, and variance

To calculate feature vectors for a pixel, we used a 9x9 estimation window (this size was also used by Cesmeli).[27] Equation (5), by multiplying the inverse of the summation of the correlation matrices (4x4) with the summation of the vectors (4x1), yields four parameters (4x1). After each pixel in the image has had its four respective parameters calculated, we calculate variance:

$$\sigma = \frac{1}{(u^2)} \sum_{r, r\pm\tau_j \in R(s)} [y_r - \hat{\Theta} Q(r)]^2 \tag{6}$$

where u is equal to the size of the estimation window. Because of our estimation window size of 9x9, u would be equal to 9.

Usually, the four GMRF parameters and variance are directly used as the feature vectors; however, as Cesmeli stated, a new set of feature vectors, taking on the property of the variance equation, are more discriminatory in detecting different textures:

$$f_j = \frac{1}{(u^2)} \sum_{r, r\pm\tau_j \in R(s)} [y_r - \hat{\theta}_j Q_j(r)]^2 \tag{7}$$

where $Q_j(r)$ is the jth component of $Q(r)$ and $j = 1, 2, 3, 4$.[27]

As one can see, Equation (7) is very similar to the variance equation (6). Because of this similarity, the new feature vectors $(f_1, f_2, f_3, \text{and} f_4)$ behave like variances in their four respective orientations: 0°, 90°, 45°, and 135°.[27] We use these four response images, along with the variance response image, as our five feature vectors for MRF in our CBIR system. (see examples in Figure 3)

Similarity Measures

Since the Haralick co-occurrence features are global, they result in a one-dimensional feature vector for each image. However, the Gabor and Markov features are local, so they result in a two-dimensional feature response for each image. Thus, we could not use the same similarity measures across all feature types.

The Haralick features were compared with three different distance measurements: Euclidean Distance, Manhattan Distance, and Chebyshev Distance.

For points $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, ..., q_n)$ where n is the number of image features:

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{8}$$

$$\text{Manhattan Distance} = \sum_{i=1}^{n} |p_i - q_i| \tag{9}$$

$$\text{Chebyshev Distance} = max_i(|p_i - q_i|) \tag{10}$$

For the Gabor and MRF feature vectors, we used a second method, in which a histogram is created from each response. The similarity between histograms can then be found using the Chi-Squared Statistic or the Jeffrey Divergence. For histograms $f(i; X)$ and $f(i; Y)$ (bin i, images X and Y) their similarity can be found by:

$$\text{Chi-Squared Statistic}(X,Y) = \sum_i \frac{(f(i;X) - \hat{f}(i))^2}{\hat{f}(i)} \tag{11}$$

$$\text{Jeffrey Divergence}(X,Y) = \sum_i f(i;X) log \frac{f(i;X)}{\hat{f}(i)} + f(i;Y) log \frac{f(i;Y)}{\hat{f}(i)} \tag{12}$$

where $\hat{f}(i) = [f(i;X) + f(i;Y)]/2$.

We have chosen these two measures since they represent two different approaches to similarity measurement. The Chi-Squared Statistic is a nonparametric similarity test and the Jeffrey divergence is a information-theory divergence. More information on these and other similarity measures can be found in the works of Rubner, Puzicha, et al.[21, 28]

LIDC Lung Nodule Project

The Lung Image Database Consortium (LIDC) maintains a database containing lung CT images and information about nodules shown in these images, including nine physician annotations regarding particular nodule features: calcification, internal structure, subtlety, lobulation, margin, sphericity, malignancy, texture and spiculation.[29]

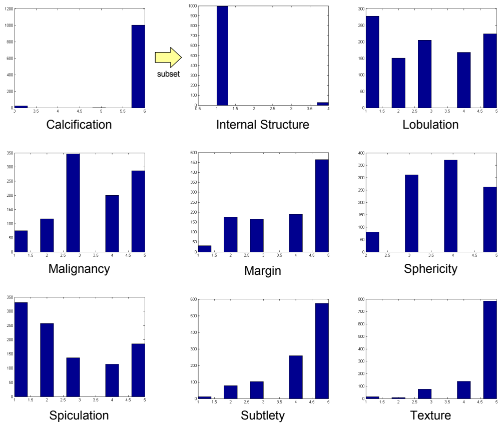


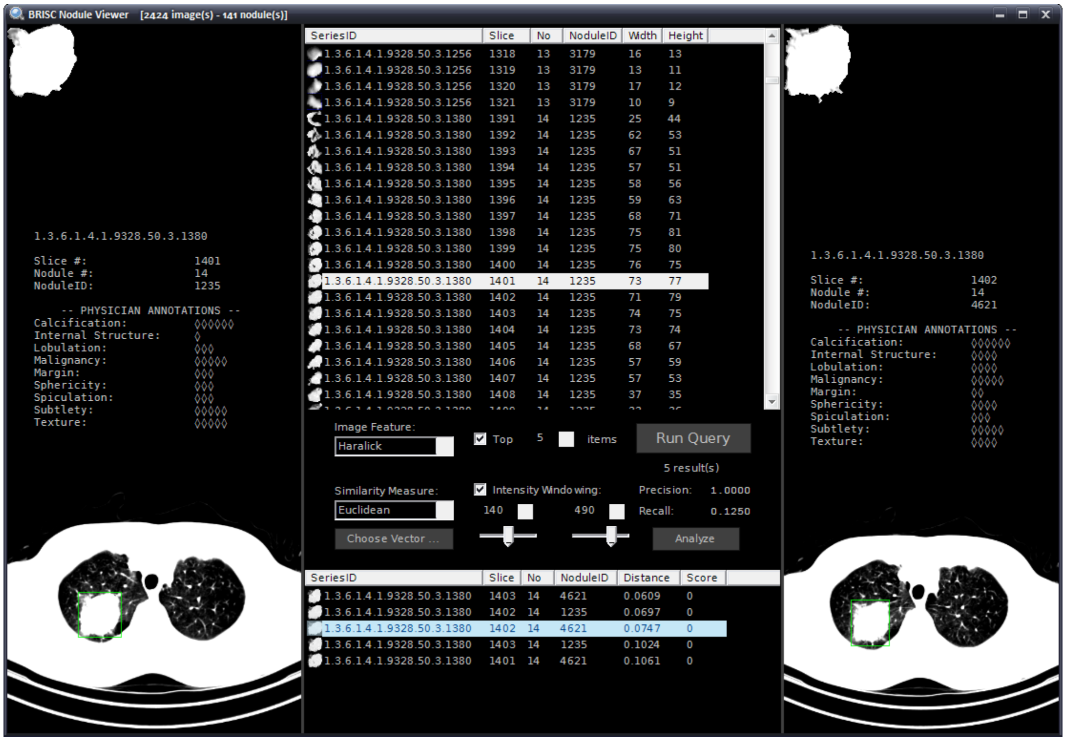
Figure 4. Annotations

All of these features are rated on an integer scale from 1 to 5 (except calcification, which is rated on a scale from 1 to 6). An examination of the feature histograms (see Figure 4) reveals that several of them (calcification, internal structure, subtlety, and texture) are almost entirely dominated by one or two major values. Thus, these particular ratings will not help much when trying to find correlations between image features and physician ratings.

The data is separated into 90 cases, each containing around 100-400 DICOM images (514KB each) and an XML data file containing the physician annotations. We extracted the XML data and used centroid calculations to determine which images are of the same nodule. Then we extracted the nodule images from the full-size CT lung scans. This produced DICOM files of the nodules, along with a collection of XML files with all of the feature data, physician annotations and metadata for each nodule image.

We discarded all nodule images smaller than 5x5 pixels (around 3x3 mm) since images this small would not yield meaningful texture data (this minimum size was also used by Kim et al.[12]) After discarding these images and ones with multiple contours, the final database contained 2424 images of 141 unique nodules. The median image size in pixels is 15x15 and the median actual size is approximately 10x10 mm. The smallest nodules are roughly 3x3 mm, while the largest are over 70x70 mm. Eighty-eight percent of the images are under 20x20 mm.

The system interface was written in C# using the .NET framework and began as a simple viewer to examine one image at a time, and was then expanded to allow side-by-side comparison of two images. Later, feature vector distance calculation was added as a way to examine the similarity of the images. The next step was to expand the program into its current state (see Figure 5): a full CBIR program that allows the user to select a query image and a threshold. The program then analyzes all the images, applies the similarity measures and determines which images are closest to the query image. It discards all images with a distance greater than the threshold value and then ranks the remaining images from closest to furthest from the query image. The interface also allows the user to choose which texture descriptors to include in the feature vector (if using Haralick features).



```
1.3.6.1.4.1.9328.50.3.1380

Slice #:                  1401
Nodule #:                 14
NoduleID:                 1235

    -- PHYSICIAN ANNOTATIONS --
Calcification:            ◊◊◊◊◊◊
Internal Structure:       ◊
Lobulation:               ◊◊◊
Malignancy:               ◊◊◊◊◊
Margin:                   ◊◊◊
Sphericity:               ◊◊◊
Spiculation:              ◊◊◊
Subtlety:                 ◊◊◊◊◊
Texture:                  ◊◊◊◊◊
```

SeriesID	Slice	No	NoduleID	Width	Height
1.3.6.1.4.1.9328.50.3.1256	1318	13	3179	16	13
1.3.6.1.4.1.9328.50.3.1256	1319	13	3179	13	11
1.3.6.1.4.1.9328.50.3.1256	1320	13	3179	17	12
1.3.6.1.4.1.9328.50.3.1256	1321	13	3179	10	9
1.3.6.1.4.1.9328.50.3.1380	1391	14	1235	25	44
1.3.6.1.4.1.9328.50.3.1380	1392	14	1235	62	53
1.3.6.1.4.1.9328.50.3.1380	1393	14	1235	67	51
1.3.6.1.4.1.9328.50.3.1380	1394	14	1235	57	51
1.3.6.1.4.1.9328.50.3.1380	1395	14	1235	58	56
1.3.6.1.4.1.9328.50.3.1380	1396	14	1235	59	63
1.3.6.1.4.1.9328.50.3.1380	1397	14	1235	68	71
1.3.6.1.4.1.9328.50.3.1380	1398	14	1235	75	81
1.3.6.1.4.1.9328.50.3.1380	1399	14	1235	75	80
1.3.6.1.4.1.9328.50.3.1380	1400	14	1235	76	75
1.3.6.1.4.1.9328.50.3.1380	1401	14	1235	73	77
1.3.6.1.4.1.9328.50.3.1380	1402	14	1235	71	79
1.3.6.1.4.1.9328.50.3.1380	1403	14	1235	74	75
1.3.6.1.4.1.9328.50.3.1380	1404	14	1235	73	74
1.3.6.1.4.1.9328.50.3.1380	1405	14	1235	68	67
1.3.6.1.4.1.9328.50.3.1380	1406	14	1235	57	59
1.3.6.1.4.1.9328.50.3.1380	1407	14	1235	57	53
1.3.6.1.4.1.9328.50.3.1380	1408	14	1235	37	35

Image Feature: Haralick

☑ Top 5 items

5 result(s)

Similarity Measure: Euclidean

☑ Intensity Windowing: 140 490

Precision: 1.0000

Recall: 0.1250

SeriesID	Slice	No	NoduleID	Distance	Score
1.3.6.1.4.1.9328.50.3.1380	1403	14	4621	0.0609	0
1.3.6.1.4.1.9328.50.3.1380	1402	14	1235	0.0697	0
1.3.6.1.4.1.9328.50.3.1380	1402	14	4621	0.0747	0
1.3.6.1.4.1.9328.50.3.1380	1403	14	1235	0.1024	0
1.3.6.1.4.1.9328.50.3.1380	1401	14	4621	0.1061	0

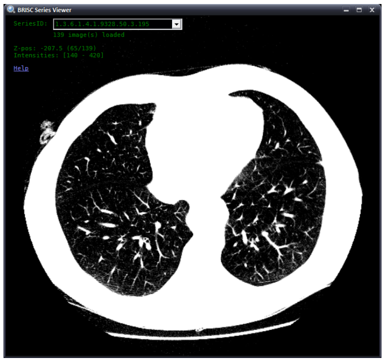
```
1.3.6.1.4.1.9328.50.3.1380

Slice #:                  1402
Nodule #:                 14
NoduleID:                 4621

    -- PHYSICIAN ANNOTATIONS --
Calcification:            ◊◊◊◊◊◊
Internal Structure:       ◊◊◊◊
Lobulation:               ◊◊◊◊
Malignancy:               ◊◊◊◊◊
Margin:                   ◊◊
Sphericity:               ◊◊◊◊
Spiculation:              ◊◊◊
Subtlety:                 ◊◊◊◊◊
Texture:                  ◊◊◊◊
```

Figure 5. Nodule Viewer

Since we needed to access the DICOM pixel data directly, we decided to use a C# DICOM library called openDICOM.net[30] available under the LGPL license to import pixel and header information from DICOM files. We also wrote a simple DICOM series viewer to explore the original lung data (see Figure 6).



```
SeriesID: 1.3.6.1.4.1.9328.50.3.195
          139 image(s) loaded

Z-pos: -207.5 (65/139)
Intensities: [140 - 420]

Help
```

Figure 6. Series Viewer

The nodule viewer is currently capable of the following:

- Importing the original, raw LIDC data into the viewer formats
- Viewing all DICOM series, with window contrast adjustment and zooming
- Viewing all nodules and their original DICOM images
- Calculating Haralick statistics, Gabor responses and Markov features on segmented DICOM images
- Nodule retrieval based on Haralick descriptors, with the option to customize the feature vector used
- Nodule retrieval based on Gabor or Markov responses
- Limit responses by number ("top N items")
- Perform on-the-fly DICOM window level contrast enhancement

This project has been released under the GNU General Public License and is available free of charge on the internet at http://brisc.sourceforge.net.

Results

Generally, CBIR systems are evaluated with respect to their performance relative to human observations, since image retrieval is only meaningful as a service to human operators.[31] Thus, the optimal performance measures are obtained when "ground truth" ratings are available for the data set. These ratings should provide a independent, objective descriptors of the data. The system can then be evaluated with respect to how many returned results have similar ratings (precision) or how many of the similarly rated images in the database were returned (recall).

As described above, our data set contained ratings of every nodule by four observing radiologists. We had expected to be able to use these ratings in the evaluation of our system, so we first tried to correlate Haralick features with physician annotations.

After performing correlation analysis between all possible Haralick feature vectors and physician annotation, the highest R-value we obtained was 0.58 for calcification and the vector: [homogeneity, cluster tendency, inverse variance]. This is fairly meaningless, however, since 98% of the nodules had the same calcification rating. The largest correlation with a well-distributed annotation was 0.25 for malignancy with the vector: [contrast, entropy, third order moment, cluster tendency].

The problem is that these annotations are very subjective, and physicians rarely agreed on nodule ratings (even of the exact same nodule). Since other systems use physician ratings in their evaluation[8],[32] it is difficult to know whether this is a general problem with medical image analysis or if the problem is specific to this database. However, if the physicians cannot agree on a common rating system, then any performance analysis using the ratings is flawed from the beginning. At this point, the problem seems to be one of ontological standardization when annotating lung nodules, which is outside the scope of our project.

Since the annotation analysis was not providing solid ground truth, we decided to base our precision and recall calculations on the idea that the first results returned by the system for a particular nodule should be other instances of that same nodule, perhaps on a different CT slice or marked and rated by a different radiologist. Thus, ground truth was determined by objective, *a priori* knowledge about the nodules. In this way, we have defined precision and recall as:

$$\text{Precision} = \frac{\#\text{ of retrieved instances of the query nodule}}{\#\text{ of retrieved images}}$$

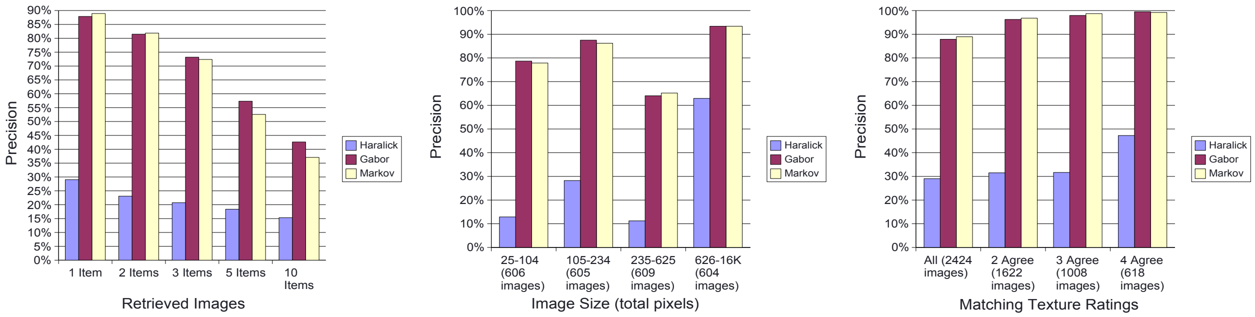
$$\text{Recall} = \frac{\#\text{ of retrieved instances of the query nodule}}{\#\text{ of total instances of the query nodule}}$$

We focus on precision scores, since in a large database, the recall is limited severely by the number of retrieved images relative to the size of the database. Thus, we did not consider recall to be a significant measure of our system's performance. This view was also taken by the ASSERT project.[8]

After running all of the preliminary trials with respect to the Haralick descriptors, we found that the worst performance was obtained when the feature vector was comprised of only one or two features. When grouped by number of retrieved images, the best mean precisions range from 20% to 29%. Comparing the results, we found that four features appear in all five of the best feature vectors: contrast, homogeneity, entropy and sum average. The best similarity measure appeared to be Manhattan, which produced four of the five best results. Thus, these four features were compared using the Manhattan distance in all further trials.

After determining which co-occurrence features to use, we ran multiple trials using co-occurrence, Gabor and Markov features to examine precision as various parameters are changed.

Figure 7(a) shows that when we vary the number of items retrieved, Gabor and Markov perform nearly identically, with the best mean precision of about 88% when one item is retrieved. Figure 7(a) also shows that Markov performs similarly to Gabor when less than five items are retrieved. However, for five and ten images retrieved Gabor shows a marked improvement over Markov. Co-occurrence matrices perform noticeably worse than both Gabor and Markov with a mean precision of only 29% when retrieving one item. One possible explanation might be that the co-occurrence model encodes the texture information at the global level while both Gabor and Markov are calculated at the pixel level.



(a) Images Retrieved

(b) Image Sizes

(c) Physician Agreement

Figure 7. Image Retrieval Results

Figure 7(b) shows what happens as precision calculations are done for different sizes of nodule images. The nodule database was divided into four equal groups based on the size of the nodule images and precision calculations were run with one item retrieved. Figure 7(b) shows that Markov and Gabor perform nearly identically and co-occurrence again performs worse. The graph also shows that all methods generally perform better on larger images, except for an unexplained decrease in precision in the third group (235-625 total pixels).

Furthermore, we ran precision calculations on nodules for which radiologists agreed on the "texture" annotation (see Figure 7(c)). When just two radiologists agreed, the average precision increased from 88% to 96% for both Gabor and Markov texture models. Once three or four radiologists agreed, the precision increased to nearly 100%.

Conclusion

We have presented a software library for content-based image retrieval of CT lung nodule images. At this point, it appears that Gabor response features outperform Haralick descriptors in improving the precision of our system. Gabor and Markov descriptors perform similarly, but Gabor features are preferred since they are quicker to calculate and compare. Unfortunately, the ratings used in lung nodule annotation do not seem to be consistent, and this poses an unsolved problem for content-based image retrieval evaluation.

Future Work

We expect that Haralick would give better results if applied locally instead of globally, as shown in the work done in 1998 by Shyu et al.[7] In addition, the Gabor transformation process contains several opportunities for optimization, and Markov might see an improvement with the addition of noise suppression. Our system could also be improved by introducing a "customized-queries" approach (CQA), which divides images into subcategories before applying similarity measures to the image descriptors, a method that has been shown to be effective for high resolution CT lung images.[33] There also exists the possibility of using the various types of texture models together, or combining our content-based algorithms with semantic content- or metadata-based retrieval algorithms for greater precision.[34] Finally, we plan to provide support for the integration of our system into the radiologist workstation project at Northwestern Memorial Hospital.

REFERENCES

1. *Cancer Facts and Figures*, American Cancer Society, 2006.
2. C. I. Henschke, D. I. McCauley, D. F. Yankelevitz, D. P. Naidich, G. McGuinness, O. S. Miettinen, D. M. Libby, M. W. Pasmantier, J. Koizumi, N. K. Altorki, and J. P. Smith, "Early lung cancer action project: overall design and findings from baseline screening," *The Lancet* **354**, pp. 99–105, July 1999.
3. Y. Kawata, N. Niki, H. Ohmatsu, M. Kusumoto, R. Kakinuma, K. Yamada, K. Mori, H. Nishiyama, K. Eguchi, M. Kaneko, and N. Moriyama, "Pulmonary nodule classification based on nodule retrieval from 3-d thoracic ct image database," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2004)*, 2004.
4. P. P. Ohanian and R. C. Dubest, "Performance evaluation for four classes of textural features," *Pattern Recognition* **25**(8), p. 819, 1992.
5. T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: a quantitative comparison," in *DAGM 2004, Pattern Recognition, 26th DAGM Symposium*, pp. 228–236, 2004.
6. H. Mller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications - clinical benefits and future directions," *International Journal of Medical Informatics* **73**, pp. 1–23, February 2004.
7. C. Shyu, C. Brodley, A. Kak, A. Kosaka, A. Aisen, and L. Broderick, "Local versus global features for content-based image retrieval," in *IEEE Workshop on Content-Based Access of Image and Video Libraries*, June 1998.
8. C.-R. Shyu, C. Brodley, A. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick, "Assert: A physician-in-the-loop content-based retrieval system for hrct image databases," *Computer Vision and Image Understanding* **75**, pp. 111–132, July/August 1999.
9. A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C.-R. Shyu, and A. Marchiori, "Automated storage and retrieval of thin-section ct images to assist diagnosis: System description and preliminary assessment," *Radiology* **228**, pp. 265–270, July 2003.
10. C.-T. Liu, P.-L. Tai, A. Y.-J. Chen, C.-H. Peng, T. Lee, and J.-S. Wang, "A content-based ct lung image retrieval system for assisting differential diagnosis images collection," in *2001 IEEE International Conference on Multimedia and Expo (ICME'01)*, 2001.
11. F. Chabat, G.-Z. Yang, and D. M. Hansell, "Obstructive lung diseases: Texture classification for differentiation at ct," *Radiology* **228**, pp. 871–877, September 2003.
12. D.-Y. Kim, J.-H. Kim, S.-M. Noh, and J.-W. Park, "Pulmonary nodule detection using chest ct images," *Acta Radiologica* (44), pp. 252–257, 2003.
13. Y. Matsuki, K. Nakamura, H. Watanabe, T. Aoki, H. Nakata, S. Katsuragawa, and K. Doi, "Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution ct," *American Journal of Roentgenology* , pp. 657–663, March 2002.
14. R. Garnavi, A. Baraani-Dastjerdi, H. A. Moghaddam, M. Giti, and A. A. Rad, "A new segmentation method for lung hrct images," in *Proceedings of the Digital Imaging Computing: Techniques and Applications*, 2005.
15. A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, pp. 1349–1380, December 2000.
16. S. Antani, L. R. Long, and G. R. Thoma, "Content-based image retrieval for large biomedical image archives," in *Proceedings of 11th World Congress on Medical Informatics (MEDINFO) 2004*, September 2004.
17. A. Materka and M. Strzelecki, "Texture analysis methods - a review," tech. rep., Technical University of Lodz, Institute of Electronics, 1998. COST B11 report.
18. R. Susomboon, D. Raicu, and J. Furst, "Pixel-based texture classification of tissues in computed tomography," in *CTI Research Symposium*, April 2006.
19. T. Galatard, J. Montagnat, and I. E. Magnin, "Texture based medical image indexing and retrieval: application to cardiac imaging," *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval* , pp. 135–142, 2004.

20. T. Andrysiak and M. Choras, "Image retrieval based on hierarchical gabor filters," *International Journal Applied Computer Science* **15**(4), pp. 471–480, 2005.
21. J. Puzicha, T. Hofmann, and J. M. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, 1997.
22. D. A. Clausi and M. E. Jernigan, "Designing gabor filters for optimal texture separability," *Pattern Recognition* **33**, pp. 1835–1849, 2000.
23. C. Bouman, "Markov random fields and stochastic image models," in *1995 IEEE International Conference on Image Processing*, 1995. Tutorial notes.
24. C. Chen, L. Pau, and P. W. (eds.), *The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, World Scientific Publishing Company, 1998.
25. S. Li, *Markov Random Field Modeling in Computer Vision*, Springer-Verlag, 1995.
26. M. Goktepe, N. Yalabik, and A. Volkan, "Unsupervised segmentation of gray level markov model textures: Hierarchical self organizing maps," in *Proceedings of the 1996 International Conference on Pattern Recognition*, pp. 90–94, 1996.
27. E. Cesmeli and D.Wang, "Texture segmentation using gaussian-markov random fields and neural oscillator networks," *IEEE Transactions on Neural Networks* **12**, pp. 394–404, March 2001.
28. J. Puzicha, Y. Rubner, C. Tomasi, and J. M. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," in *ICCV (2)*, pp. 1165–1172, 1999.
29. *LIDC Lung Nodule Image Database*, National Cancer Imaging Archive (https://imaging.nci.nih.gov/ncia/).
30. *openDICOM.net*, SourceForge (http://opendicom.sourceforge.net/).
31. N. V. Shirahatti and K. Barnard, "Evaluating image retrieval," in *Proceedings of the 2005 IEEE Computer Science Conference on Computer Vision and Pattern Recognition*, 2005.
32. C.-R. Shyu, A. Kak, C. E. Brodley, and L. S. Broderick, "Testing for human perceptual categories in a physician-in-the-loop cbir system for medical imagery," in *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1999.
33. J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen, "Unsupervised feature selection applied to content-based retrieval of lung images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(3), pp. 373–378, 2003.
34. S. Atnafu, R. Chbeir, and L. Brunie, "Content-based and metadata retrieval in medical image database," in *Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems*, 2002.