Task-Oriented Medical Image Retrieval

William Horsthemke, Daniela Raicu, and Jacob Furst

DePaul University, Chicago, IL, 60604, USA
horsthemke@acm.org, {draicu, jfurst}@cs.depaul.edu

Abstract. Many clinical tasks depend upon the proper interpretation of medical images and will benefit from access and reference to similar, relevant images. In this paper, we present CBIR approaches to two clinical tasks and discuss their specific challenges. The first CBIR system retrieves anatomical regions in Computed Tomography (CT) studies of the chest and abdomen. The system can be used to provide *context-sensitive tools* for computer-aided diagnosis (for example, apply lung CAD algorithms on a region of interest only if the anatomical structure was identified as lung). The second CBIR system retrieves pathologies specific to an anatomical structure, such as nodules present in the CT studies of the lung. This system can be used directly as a computer-aided diagnosis system for case-based and evidence-based medicine. Both systems are evaluated using texture image features and several similarity measures. Given the fact that finding similar pathologies for an anatomical structure is a more difficult problem than finding similar anatomical structures, the second system is also evaluated with respect to the radiologists' variability in the process of pathology interpretation. We found that the retrieval precision improved from 88% to 96% and 100% when the ground truth data was verified by two (96%) or three (100%) radiologists.

1 Introduction

Content-based image retrieval (CBIR) has emerged as an index and query method to retrieve and use the abundance of useful medical images, often available only through a limited set of textual annotations recorded about the patient or procedure with sparse information about the image itself. Education and training will benefit from access to patient cases similar to a given image. Reference to cases with similar features and known pathologies will assist and support medical decision-making as in case based reasoning [15] or evidence-based medicine [7]. Computer-aided diagnosis benefits from automated access to similar images for performance analysis or functional use of specific identification and similarity methodologies developed for CBIR.

In this paper, we present two CBIR systems. The first measures the image features of anatomic regions of the abdomen and chest and explores data representations and similarity measures which can express the similarity between images of the same type of anatomy. This anatomic region CBIR achieves an overall precision of almost 92%, with 100% for many specific anatomies. The second CBIR, instead of retrieving similar anatomical structures, attempts to find similar pathologies within an anatomical structure, in particular the lung. Given that the pathologies are lung nodules manually delineated and annotated by up to four radiologists within the NIH Lung Image Database Consortium (LIDC) [3], there is a lot of variability in the interpretation of a nodule image similarity across the images for the same nodule (sequence of CT slices in which the nodule appears) and across the nodules from the data set. Therefore, the precision and recall results are also evaluated taken into account the agreement among radiologists. Since the computer results are expected to match the human perception, and the proposed CBIR system uses low-level texture features, the radiologist agreement was quantified only with respect to texture, one of the nine characteristics used in the LIDC to describe the visual appearance of a nodule. Our results show that the pulmonary nodule retrieval precision improved from 88% to 96% and 100% when the ground truth data was verified by two (96%) or three (100%) radiologists.

We discuss the general requirements for CBIR systems, discuss the methodology for performance assessment, and illustrate how the combination of factors affecting overall performance presents an optimization problem addressed by parameter sweeps or sensitivity analysis.

2 Related Work

Development of a clinically useful CBIR for medical applications depends upon several key research areas including image processing, machine learning, computer-aided diagnosis, as well as previous CBIR research. Several projects have developed CBIR for general medical applications. Lehmann et al. [21] designed IRMA to automatically extract global features to classify images according to image modality, orientation, body region, and target organ, using this information to select and extract local features using a-priori diagnostic models and image atlases. In GMM-KL, Greenspan et al. [11] uses a probabilistic image representation to categorize, match, and retrieve images by body region. Using texture, intensity, and spatial information, GMM-KL finds similar regions through unsupervised clustering. Other general purpose CBIR projects include works by Chu et al. [9] on KMED; by El-Kwae [20] on COBRA; and by Muller et al. [24] on MedGIFT.

Most medical CBIR applications focus on particular organs, anatomical structures, modalities, or diagnostic categories, such as cardiology, pathology, and radiology. Glatard et al. [11] used Gabor texture filters [1] to automatically identify, segment, and retrieve images with similar myocardial contraction from cardiac Magnetic Resonance Imaging (MRI). For retrieval of similar microscopic pathology images, Zheng et al. [32] extracts color histograms, texture, and fourier and wavelet coefficients and compares images using the vector dot product as a similarity metric. Antani et al. [2] developed shape-based methods to index and retrieve over 17,000 spinal images to allow for query by image or sketch. Korn et al. [19] used shape as a feature to index and retrieve tumors identified with mammography. Trourassi et al. [30] studied the effect of entropy-based information-theoretic similarity measures on the retrieval performance of mammographic masses. Dy et al. [10] describe an unsupervised method for feature selection (reduction) in ASSERT which uses various features such as co-occurrence textures, Fourier descriptors, and moments extracted from radiologist-marked suspect regions in CT of the lung. ASSERT contains a reference database with pathology-defined ground truth for supervised classification of disease categories.

Some medical CBIR applications attempt a global understanding of the image while others focus on specific anatomies or pathologies; often requiring a different definition of ground truth (relevance). Our anatomical region CBIR, as well as two global approaches (IRMA and GMM-KL), identifies specific anatomical structures and provides a context-sensitive tool for computer aided diagnosis: (for example, apply lung CAD algorithms on a region of interest only if the anatomical structure was identified as lung). For these applicatons, reference to an atlas-based anatomical ground truth might suffice, with relevance defined as correct anatomical identification.

Our pathology-based lung nodule CBIR supports computer-aided diagnosis for case-based and evidence-based medicine and requires ground truth for both the nodule diagnosis (a binary, presence or absence of disease) and radiologists' assessments of various diagnostic characteristics. Ground truth for the primary diagnosis presents challenges, often not fully appreciated or addressed by other applications. In many systems, ground truth is decided by a single, project radiologist, though often confirmed by pathology, histology, or disease progression. However, in many diagnostic applications, especially early stage lung nodule diagnosis, significant disagreement among expert radiologists is expected [3]. Ground truth about radiologists' assessments of various diagnositic characteristics presents an additional challenge. Only one other study (ASSERT) attempted to interpret a radiologist description of an image-based diagnostic characteristic (homogeneity). Studies of inter-observer variablity confirm the challenge faced by automated methods. In a study in ultrasound breast imaging, Baker et al. [4] found a lack of uniformity among radiologists' use of descriptive terms and inconsistent diagnoses, even though they used a benchmark lexicon proposed for describing the appearance of masses.

There are not many studies which analyze the direct correlation between the computer results and the radiologists' perceptual similarity for a pair of images. Recently, Li et al [16] studied correlation between the similarity of computer extracted image features and radiologists' visual similarity perception using an inter-observer study with ten radiologists. They combined multiple image features using a neural network to predict a reliable psychophysical similarity measure for selection of similar images for malignant and benign nodules. Raicu et al. [26] used multiple linear regression to predict the scoring by multiple radiologists of diagnostic characteristics of lung nodules using shape, size, intensity, and texture features. Recent work by Muramatsu et al. [22] shows that averaging the readings from the same observers will provide a reliable method for establishing ground truth for clustered microcalcifications in mammography, even when there exists large intra- and inter-observer variability.

3 CBIR Framework

Many CBIR system share a common framework to fulfill the basic requirements of image comparison and retrieval, including the two discussed herein. The fundamental requirement is a method to extract salient features from the image, either from the entire image or regions of interest within the image. In our two systems, a region of interest is selected and segmented before feature extraction. Other applications operate on entire images for analysis and comparison. Medical applications tend to focus on regions of interest. Figure 1 illustrates an example of the process flow from image to retrieval of similar images using the process model of the anatomical structure image retrieval as an example.

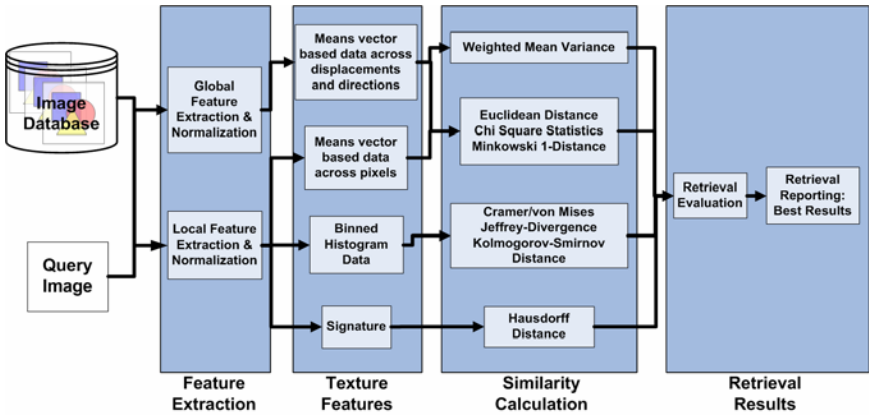
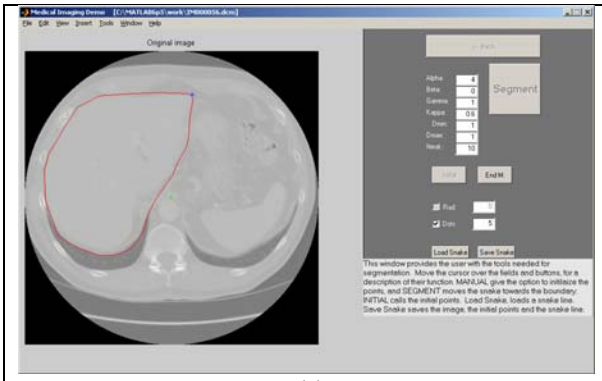
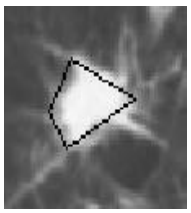
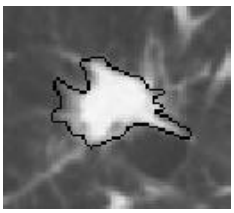


Fig. 1: Example CBIR Process

3.1 Region of Interest

Identification and segmentation pose a significant challenge to CBIR. In many applications, manual intervention is required either to locate and identify a region of interest or to perform the actual segmentation. Much work on automatic location and segmentation has been performed in the computer-aided diagnosis community but few CBIR systems have incorporated these techniques. For organ identification, our system includes a semi-automated segmentation tool, requiring only the manual selection of a few points around the region of interest as shown in the Table 1(a). For pulmonary nodule diagnosis, the datasets include one or more radiologist-drawn outlines for use as segmentation templates. Table 1(b, c) also illustrates the significant variability in radiologist-drawn outlines. The images contain outlines of the same pulmonary nodule (same CT slice) drawn by two radiologists, and differ significantly in area included as well as level of detail about the boundary of the nodule: a significant clinical, diagnostic feature.

Table 1. Semi-automated segmentation of anatomical regions and illustration of radiologist-drawn outlines

 (a)	 (b)	 (c)

3.2 Feature Extraction

Selection of features remains an important challenge and our choice of texture-based features is based on our earlier nodule characteristics prediction [26] and organ classification research [27]. Muller et al. [23] compared several texture analysis methods and color (grey level) quantization. Other useful features to extract include image intensity, region size and shape, and statistical moments. Our choice for texture measurements range from statistical (Haralick statistics of co-occurrence texture matrices), filters (Gabor transforms), to model-based approaches (Markov Random Fields (MRF)). Co-occurrence is applied for both CBIR systems while the Gabor and MRF are applied only on pulmonary nodules.

Haralick texture descriptors capture various spatial dependencies and structures within an image [13]. Due to redundancy and correlation, only ten Haralick descriptors are recommended: Entropy, Energy, Contrast, Homogeneity, SumMean, Variance, Maximum Probability, Inverse Difference Moment, Cluster Tendency, and Correlation. The Haralick statistics are computed from co-occurrence matrices which represent the conditional joint probability of gray level pixel-pairs based upon their separation (displacement) and orientation (direction) [13].

Various methods of computing the co-occurrence matrix are available. One method considers all pairs of pixels (often within a small neighborhood) while another considers only a fixed set of directions and displacements. Often four directions (0^0, 45^0, 90^0, and 135^0) are used and N-1 displacements where N is the size of the neighborhood window. For global texture measurements our implementation computes co-occurrence matrices in four directions (0^0, 45^0, 90^0, and 135^0) and five displacements (1, 2, 3, 4, and 5) , generating twenty matrices for each segmented image. From each of these twenty matrices, ten Haralick descriptors are computed, averaged, and recorded as a mean-based feature vector for the segmented image.

For pixel-level texture measurements, all-pairs co-occurrence matrices are computed for a 5 by 5 window, small enough to capture local texture properties while large enough to capture enough samples for statistical significance. Using this approach, a co-occurrence matrix is created for each pixel rather than one for each choice of distance and direction and ten Haralick feature descriptors are calculated. Using pixel-level co-occurrence matrices, we derive three representations: 1) mean vector-based data, 2) binned histogram data and 3) texture signatures.

The mean vector-based data representation consists of the average of the set of vectors, such as Haralick texture descriptors. The binned histogram data representation consists of texture values grouped within equal-width bins. The number of bins and their placement are important parameters as they determine how crudely, or how well, the underlying probability distribution (obtained by quantizing the responses into bins and normalizing such that the sum over all bins is unity) is modeled: too many bins will overfit the data and introduce noise while too few bins will make the binning crude. In our experimental results, a number of 256 equal-size bins produced the best results for this representation.

The texture signature representation uses the k-d tree clustering algorithm [5] to permit mixed-size clusters, thus avoiding constraint of binned histograms. Proportional cluster size and intra-cluster variance, as a percentage of parent size and variance, respectively, were chosen as stopping criteria. Experimental sensitivity analysis of varying these two criteria to retrieval performance discovered optimal performance using 10% variance and 20% cluster size, measured using directed Hausdorff distance [31].

Gabor filters [1] capture localized texture frequencies by convolving an image with Gaussian modulated frequency transform. Our method uses the resulting means and standard deviations of the twelve (12) responses from applying four (4) orientations (0º, 45º, 90º, and 135º) and three (3) frequencies to a 9x9 image window.

Markov Random Fields (MRFs) model the local contextual information of an image [6] using a window approach. Using the Cesmeli [8] algorithm, we convolved the image with a 9x9 MRF window and computed the mean response for 4 angular rotations (0°, 45°, 90°, 135°) and variance, for a total of 5 MRF features.

3.3 Similarity Measurement

3.3.1 Definitions

Comparing images or regions of interest becomes the next major challenge and several similarity measures have been identified, some applicable to only certain types of feature representation. Many applications can

use standard Euclidean distance for similarity but often additional measurements are indicated. For histogram-based data, Rubner et al. [28] defines four categories of similarity measurements: heuristic distances, non-parametric test statistics, information theory divergences, and ground distances using the following notation:

$$Feature\ Vectors:\ H = (h_1, h_2, \ldots, h_{n)}, K = (j_1, k_2, \ldots, k_n), \text{mean of } x = \mu_X, \text{stddev of } x = \sigma_X,$$

$f^i(j;H)$ is binned histogram and $F^i(j;H)$ is cumulative histogram of feature i across j bins

Heuristic distance metrics: 1) Minkowski 1-distance, d_{Lr} (city block distance or L_1 norm) (Equation (1)), 2) weighted-mean-variance, d_{wmv} (uses the means and standard deviations for each of the considered features, Equation (2)). For comparison purposes we make use of the Minkowski distance using both local and global level vector based data. When r is equal to 2 the Minkowski distance becomes the Euclidean distance. When r is equal to 1, the Minkowski 1-distance is known as the Manhattan distance. For the local representation, the histogram representation still allow for vector-based representation by comparing the number of elements within each histogram bin.

$$d_{L_r}(H,K) = \left(\sum_i |h_i - k_i|^r\right)^{\frac{1}{r}} \quad \textbf{(1)}$$

$$d_{wmv}(H,K) = \sum_i \frac{|\mu_i(H) - \mu_i(K)|}{|\sigma(\mu_i)|} + \frac{|\sigma_i(H) - \sigma_i(K)|}{|\sigma(\sigma_i)|} \quad \textit{where } i = \textit{feature index} \quad \textbf{(2)}$$

Non-parametric test statistics: 1) Cramer-von Mises, d_{CvM} (similar to the squared Euclidean distance but calculated between the distributions and as the maximal discrepancy between the cumulative distributions) (Equation (13)). Cramer-von Mises is defined as the squared Euclidean distance between the distributions and as the maximal discrepancy between the cumulative distributions. It has a desirable property to be invariant to arbitrary monotonic feature transformations. It is also a single feature calculation that does not take into account information that may exist across features [28]. For our purposes, d_{CvM} is applied to the cumulative distribution, F, for each feature, i, across j bins:

$$d_{CvM}(H,K) = \sum_i \sum_j \left(F^i(j;H) - F^i(j;K)\right)^2 \quad \textbf{(3)}$$

2) Kolmogorov-Smirnov distance, d_{KS} (used for unbinned data distributions and it is invariant to arbitrary monotonic transformations) (Equation (4)). Kolmogorov-Smirnov (d_{KS}) is normally defined only for one dimension. It is a single feature calculation that does not take into account information that may exist across features. It is a common statistical tool for unbinned distributions and it has a desirable property to be invariant to arbitrary monotonic feature transformations [28]. For our purposes, d_{KS} is applied to each feature, i, using the cumulative binned histogram data, F, where j represents the bins from H and K respectively. The Kolmogorov-Smirnov distance is then summed across all features. d_{JD} is the statistical dissimilarity metric that can be used to compute the distance between class distributions of two values of the same feature:

$$d_{KS}(H,K) = \sum_i \max_j \left(\left| F^i(j;H) - F^i(j;K) \right| \right) \quad \textbf{(4)}$$

3) Chi-square statistics, $d\chi^2$ (used to distinguish whether distributions of the descriptors differ from each other) (Equation (5)) Chi-square Statistics, $d\chi^2$, are used to distinguish whether distributions of variables differ from each other and measures how unlikely it is that one distribution was drawn from the population represented by the other [29]. Chi-square Statistics take into account the correspondence between bins with the same index. The similarity is calculated using both the local and global vector based data. So, h_i and k_i, where i goes from 1 to 10, correspond to the means of the 10 individual Haralick descriptors calculated for the given image. In our research, we also analyzed the weighted-mean-variance metric found to outperform

several metrics in the case of Gabor texture feature representation [28]. In the formula for this metric, $\sigma(\bullet)$ denotes an estimate of the standard deviation of the respective entity and $\sigma_i(\bullet)$ represents the estimate of the standard deviation for feature i given by the i^{th} Haralick texture descriptor.

$$d_{\chi^2}(H,K) = \sum_i \frac{(h_i - m_i)^2}{m_i}, \text{where } m_i = \frac{h_i + k_i}{2} \text{ is the mean histogram} \quad (5)$$

Information Theory Divergences: 1) Jeffrey-Divergence, d_{JD} (used to compute the distance between class distributions of two values of the same feature) (Equation (6)). Jeffrey-Divergence is symmetric and numerically stable when comparing two distributions [28]. It is a single feature calculation that does not take into account information that may exist across features. In this case the d_{JD} is calculated using the binned histogram data, *f,* across *j* bins and is then summed across the *i* features. 2) Kullback-Leibler (KL) divergence, d_{KL} (Equation (7)).

$$d_{JD}(H,K) = \sum_i \sum_j \left(f(j;H) \log \frac{f(j;H)}{m_j} + f(j;K) \log \frac{f(j;K)}{m_j} \right) \quad (6)$$

$$\text{where } m_j = \frac{f(j;H) + f(j;K)}{2}$$

$$d_{KL}(H,K) = \sum_i f(j;H) \log \frac{f(j;H)}{f(j;K)} \quad (7)$$

In addition to the above measures, two others are implemented as required by the different texture feature representations: Euclidean distance d_A (Equation (8)) and Hausdorff distance, d_{HD} (used for texture signature representation) (Equation (9)).The Hausdorff distance is applied to calculate the similarity among two images for the texture signature representation. A directed form of the Hausdorff distance is used because we are primarily interested in finding an image similar to our query image, *H.* This directed form calculates the minimum distance *d* from all distances between all the points in the initial set or a given image, *K*, and any point within the query image, *H* [31].

:

$$d_A(H,K) = \sum_i \sqrt{(h_i - k_i)^2} \quad (8)$$

$$d_{HD}(H,K) = \max_{h \in H} (\min_{k \in K} (\| h - k \|)) \quad (9)$$

3.4 Performance Evaluation

Medical image retrieval systems often neglect to report an overall evaluation of their retrieval performance [25]. For CBIR systems, this paper measures overall retrieval performance using precision and recall as defined by Equations (10) and (11). The method for evaluating performance interatively considers each image as query, measures each retrieval outcome, and reports overall average. Often the database is partitioned to study the effect of different image characteristics, such as nodule size or level of radiologist agreement.

$$\text{Precision} = \frac{\#of\_relevant\_retrieved\_images}{total\#of\_retrieved\_images} \quad (10)$$

$$\text{Recall} = \frac{\#of\_relevant\_retrieved\_images}{total\_\#of\_relevant\_images} \quad (11)$$

The pulmonary nodule CBIR system defines as a correct retrieved image another slice from the same nodule as the query image. For the organ retrieval CBIR, a correct retrieved image is defined as another image having the same organ label as the query image.

4 Anatomical Region CBIR

4.1 Data Acquisition and Image Segmentation

The anatomical regions were obtained and labeled through a semi-automated segmentation of 344 de-identified datasets from normal abdominal and chest CT studies. Using the Active Contour Models (ACM) algorithm [17] to allow for complex shapes, we segmented five regions: heart and great vessels, liver, renal and splenic parenchyma, and backbone. Both global and local Haralick texture features are extracted and compared since size of texture operator windows can affect classifier performance [14].

4.2 Methodology and Performance Analysis

Each of the 344 organ images serves as a query into the database containing 5 types of organs. Matching organ type is the ground truth for computing the relevance of the retrieval. The results are weighted by the percentage of organ types within the database to assess both organ type and overall performance. The reported results use k=6 for the most retrieved similar images with a recall of 8% and a precision of 92%.

This organ CBIR determined the best combination of texture feature representation and corresponding similarity measure using both pixel-level and global-level data. All eleven combinations of feature sets and similarity measures exceeded an overall precision of 80% (Table 2). Overall and individually per region, global-level outperformed pixel-level using vector representation by about 6%. Pixel-level binned histogram with Jeffrey-Divergence similarity performed the best with an overall precision of almost 92% and an individual precision of no less than 75% for the spleen which is physically similar and mistaken for the liver: backbone (100%), heart (89.7%), kidneys (96%), liver (77.87%) and spleen (75.83%).

Table 2. Precision at the global and local-level; overall performance is the weighted average

GLOBAL LEVEL VECTOR BASED PRECISION						
	Backbone	**Heart**	**Kidney**	**Liver**	**Spleen**	**OVERALL**
Euclid Distance	100.0%	90.4%	93.8%	67.8%	62.1%	87.7%
Chi Square Statistics	100.0%	90.7%	93.8%	62.9%	57.5%	86.4%
Minkowski 1 Distance	100.0%	90.1%	92.9%	69.0%	62.5%	87.8%
PIXEL LEVEL VECTOR BASED PRECISION						
	Backbone	**Heart**	**Kidney**	**Liver**	**Spleen**	**OVERALL**
Euclid Distance	100.0%	76.0%	85.8%	59.8%	46.7%	81.2%
Chi Square Statistics	100.0%	81.1%	87.7%	60.1%	47.5%	82.4%
Minkowski 1 Distance	100.0%	74.4%	85.2%	59.5%	48.8%	81.0%
Weighted-Mean-Variance	100.0%	87.2%	91.7%	58.9%	53.8%	84.5%
PIXEL LEVEL BINNED HISTOGRAM BASED PRECISION						
	Backbone	**Heart**	**Kidney**	**Liver**	**Spleen**	**OVERALL**
Cramer/von Mises	100.0%	88.8%	83.6%	64.1%	51.3%	84.0%
Jeffrey-Divergence	100.0%	91.7%	96.0%	77.9%	75.8%	91.6%
Kolmogorov-Smirnov Distance	100.0%	89.1%	89.8%	69.8%	60.0%	87.0%
PIXEL LEVEL SIGNATURE BASED PRECISION						
	Backbone	**Heart**	**Kidney**	**Liver**	**Spleen**	**OVERALL**
Hausdorff 10% variance & 20% cluster size	100.0%	81.1%	86.4%	57.8%	42.1%	81.2%

5 Pulmonary Nodule CBIR

5.1 Data Acquisition and Image Segmentation

The pulmonary nodule images are obtained from the LIDC [3] and represent lesions between 3 and 30 mm, identified by at least one of four radiologists. In addition to detecting these nodules, the radiologist outlines the nodule on each CT slice and rates their interpretation of 13 nodule characteristics (focal abnormality, complexity, convexity, radiographic opacity, calcification, internal structure, subtlety, lobulation, margin, sphericity, malignancy, texture, and spiculation). Our own work [26,27] motivated the choice of texture features.

The results were obtained on a subset of the LIDC image data. We used centroid calculations to determine which images are of the same nodule and segmented the nodule images from the full-size CT lung scans using the radiologist-drawn outlines as templates. Nodule images smaller than 5x5 pixels (around 3x3 mm) were discarded since they would not yield meaningful texture data [18]. After discarding these images and ones with multiple contours by the same radiologist, the final database contained 2424 images of 141 unique nodules. The median image size in pixels was 15x15 and the median actual size was approximately 10x10 mm. The smallest nodules were roughly 3x3 mm, while the largest were over 70x70 mm. Eighty-eight percent of the images were less than 20x20 mm.

5.2 Methodology and Performance Analysis

Each nodule image serves as a query and performance was calculated as the average precision at n where n ranged from 1 to 10. Given the significant disagreement among radiologists about the existence and scoring of nodule characteristics, a separate study on retrieval performance with varying levels of radiologist agreement was performed. In this study (along with a study of the effect of nodule size), the performance is reported as the average precision when retrieving only the top image (n = 1). Recall performance ranged from 2 to 14% depending upon the number of slices of a nodule and agreement of radiologists. Correctly retrieved images are other images from other CT slices of same nodule.

Finding the optimal retrieval performance for the LIDC pulmonary nodule database required a series of sensitivity analysis of the combination of texture descriptors, similarity measures, and number of retrieved images. Initial experiments winnowed to five the number of Haralick descriptors (contrast, homogeneity, entropy, and sum average) and to the Manhattan distance. Final performance assessment then compares the five Haralick co-occurrence descriptors using Manhattan similarity with Gabor and MRF using histogram-based distances.

As illustrated in Figures 2, 3, and 4, Gabor and MRF outperform Haralick in the precision of retrieved nodules. Figure 2 compares precision performance to the number of retrieved images which shows that Gabor and Markov retrieves a relevant image in almost 90% of the cases when only one image is requested. Since Haralick texture encodes global texture information, while both Gabor and MRF operate locally, the effect of nodule size on performance was investigated as shown in Figure 3 which shows that the difference between the global Haralick and localized Gabor and MRF decreases as the size of the nodule increases, except for an unexplained overall decrease in precision in the third group (235-625 total pixels). Inter-observer differences in pulmonary nodule radiology have been shown in nodule detection, diagnosis, and interpretation of descriptive features such as texture. To investigate the effect of radiologist variability on nodule retrieval, we performed a sensitivity analysis by varying the level of radiologist agreement, see Figure 4. When just two radiologists agreed on "texture", the average precision increased from 88% to 96% for both Gabor and Markov texture models and reached nearly 100% when three or four agreed. This illustrates the significant challenge on both CBIR and computer aided diagnosis (CAD) from radiologists' variability in establishing relevance in CBIR and ground truth in CAD.

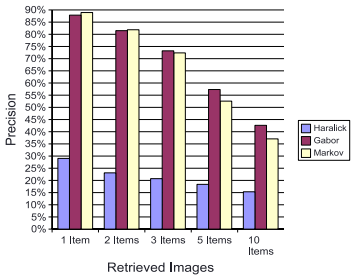


Fig. 2. Precision vs. # of retrieved images

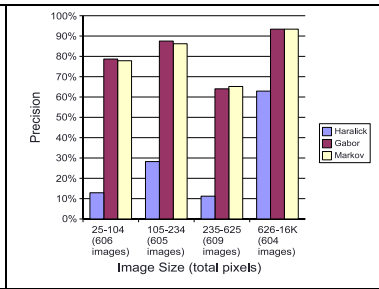


Fig. 3. Precision vs. image size

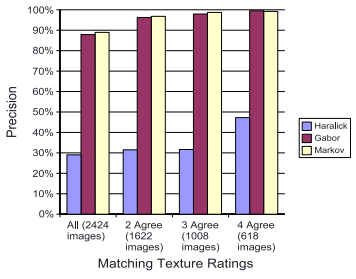


Fig. 4. Precision vs. radiologists' agreement

6 Conclusion

Overall, these initial CBIR prototypes perform well and illustrate the success and challenges faced in the enhancement and integration of such systems into clinical practice. Though we did not address the human factors and user interface in this discussion, integration and usability of these systems presents another major challenge, partially addressed in the pulmonary nodule CBIR. The anatomical region CBIR faced a challenge discriminating between liver and spleen and has subsequently been improved by the incorporation of other texture features, such as Gabor and Markov Random Fields. However, their biological and visual similarity illustrates one of the challenges in designing discriminating features and similarity measures. The distance between backbones and livers is much greater than livers to spleens and the measurement tools might need non-linear support for variable precision. This is one area where design of CBIR can be augmented from machine learning technologies.

As shown in the pulmonary nodule CBIR, the variability in ground truth among radiologists presents a significant challenge. This was illustrated by the positive relationship between retrieval performance and radiologist agreement. Typically, CBIR attempts to map image features to well-known visual appearances, but predicting varied opinions will require much more effort. However, one application of medical CBIR is on education and training where access to machine annotated images may increase the level of agreement among newly trained radiologists. Future work will include the incorporation of other feature measurements, including shape and expand the evaluation of performance to include other diagnostic characteristics such as spiculation, lobulation, and margin. Incorporating predicted diagnostic characteristics to the feature set through machine learning is another intended area of research.

References

[1] T. Andrysiak and M. Choras, "Image retrieval based on hierarchical Gabor filters," International Journal Applied Computer Science, vol. 15, no. 4, 471-480, 2005.

[2] Antani S, Long LR, Thoma G. "Content-Based Image Retrieval for Large Biomedical Image Archives" Proceedings of 11th World Congress on Medical Informatics (MEDINFO) 2004 Imaging Informatics. September 7-11 2004; San Francisco, CA, USA. 829-33.

[3] S. G. Armato, G. McLennan, M. F. McNitt-Gray, C. R.Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, A. P. Reeves, B. Y. Croft, and L. P. Clarke, "Lung Image Database Consortium: Developing a resource for the medical imaging research community," Radiology, vol. 232, no. 3, pp. 739-748, 2004.

[4] J. A. Baker, P.J. Kornguth, M.S. Soo, R. Walsh, P. Mengoni. Sonography of solid breast lesions: observer variability of lesion description and assessment. AJR 1999, 172:1621-1625.

[5] J. L. Bentley. Multidimensional binary search trees used for associative searching. Communications of the ACM, 18:509-517, 1975.

[6] C. Bouman, "Markov random fields and stochastic image models," in 1995 IEEE International Conference on Image Processing, 1995. Tutorial notes.

[7] A. A. T. Bui, R. K. Taira, J. D. N. Dionision, D. R. Aberle, S. El-Saden, H. Kangarloo, Evidence-based radiology, Academic Radiology 9 (6) (2002) 662-669.

[8] E. Cesmeli and D. Wang, "Texture segmentation using gaussian-markov random fields and neural oscillator networks," IEEE Transactions on Neural Networks, vol 12, pp. 394-404, March 2001.
[9] W. W. Chu, C. C. Hsu, A. F. Cardenas, and R. K. Taira, "Knowledgebased image retrieval with spatial and temporal constructs," IEEE Transactions on Knowledge and Data Engineering, vol. 10, no. 6, pp. 872-888, Nov. 1998.
[10] 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, vol. 25, no.3, pp. 373-378, Mar. 2003.
[11] T. Glatard, J. Montagnat, and I.E. Magnin. Texture based medical image indexing and retrieval: application to cardiac imaging. ACM SIGMM international workshop on Multimedia Information Retrieval (MIR'04), Proceedings of ACM Multimedia 2004, New-York, USA, October 15-16, 2004.
[12] Greenspan H, Pinhas AT. "Medical image categorization and retrieval for PACS using the GMM-KL framework." IEEE Trans Inf Technol Biomed. 2007 Mar;11(2):190-202.
[13] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," IEEE Trans. On Systems, Man, and Cybernetics, vol. 3, no. 6, 610-621, 1973.
[14] W. H. Horsthemke, J. D. Furst, and D. S. Raicu. "Texture Classifier Robustness for Sub-organ Sized Windows". DePaul CTI Research Symposium, 2004
[15] C. LeBozec, M.-C. Jaulent, E. Zapletal, P. Degoulet, Unified modeling language and design of a case- based retrieval system in medical imaging, in: Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA), Nashville, TN,USA, 1998.
[16] Li Q, Li F, Shiraishi J, Katsuragawa S, Sone S, Doi K: Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules. In: Medical Physics. 30(10):2584-93, 2003.
[17] Kass, M., Witkin, A., Terzopoulos, D. (1988). Snakes: Active contour models. Int'l. J. of Comp. Vis. 1(4).
[18] Kim D-Y, Kim J-H, Noh S-M, Park J-W: Pulmonary nodule detection using chest CT images. In: Acta Radiologica (44), pp. 252–257, 2003.
[19] P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas,"Fast and effective retrieval of medical tumor shapes," IEEE Transactions on Knowledge and Data Engineering, vol. 10, no. 6, pp.889-904, Nov. 1998.
[20] E. A. El-Kwae, H. Xu, and M. R. Kabuka, "Content-based retrieval in picture archiving and communication systems," Journal of Digital Imaging, vol. 13, no. 2, pp. 70-81, May, 2000.
[21] T. Lehmann, B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen. Content-based Image Retrieval in Medical Applications: A Novel Multi-step Approach. In Procs. Int. Society for Optical Engineering (SPIE), volume 3972(32), pages 312-331, February 2000.
[22] Muramatsu C., Li Q., Schmidt R., Suzuki K., Shiraishi J., Newstead G., Doi K., "Experimental determination of subjective similarity for pairs of clustered microcalcifications on mammograms: observer study results", Medical Physics. 33(9):3460-8, 2006
[23] H. Müller, A. Rosset, J-P. Vallée, A. Geissbuhler, Comparing feature sets for content-based medical information retrieval. SPIE Medical Imaging, San Diego, CA, USA, February 2004.
[24] H. Müller, P. Fabry, A. Geissbuhler. MedGIFT - Retrieving medical images by there visual content. World Summit of the Information Society, Forum Science and Society, Geneva, Switzerland, December 2003.
[25] H. Müller, N. Michoux, D. Bandon, A. Geissbuhler, A review of content-based image retrieval systems in medical applications: clinical benefits and future directions. International journal of medical informatics, Volume 73 pp 1-23.
[26] Raicu DS, Varutbangkul E, Cisneros JG, Furst JD, Channin DS, Armato III SG: Semantics and Image Content Integration for Pulmonary Nodule Interpretation in Thoracic Computed Tomography. In: SPIE Medical Imaging Conference, San Diego, CA, February 2007.
[27] D. S. Raicu, J. D. Furst, D. Channin, D. H. Xu, & A. Kurani, "A Texture Dictionary for Human Organs Tissues' Classification", Proceedings of the 8th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2004), Orlando, USA, in July 18-21, 2004.
[28] Y.. Rubner, J. Puzicha, C. Tomasi, and J.M. Buhmann. Empirical Evaluation of Dissimilarity Measures for Color and Texture. International Conference on Computer Vision-Volume 2, Corfu, Greece, September 20 - 25, 1999. pp.1165.
[29] Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's Distance as a metric for image retrieval. Technical Report STAN-CS-TN-98-86, Computer Science Department, Stanford University, September 1998.
[30] Georgia D. Tourassi, Brian Harrawood, Swatee Singh, Joseph Y. Lo, and Carey E. Floyd, "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms" Medical Physics -- January 2007 -- Volume 34, Issue 1, pp. 140-150.
[31] G. Wei, D. Li and IK Sethi. Detection of Side-view Faces in Color Images. Proceedings Fifth IEEE Workshop on Applications of Computer Vision, pp. 79-84, Palm Springs, CA, 4-6 December 2000.
[32] L. Zheng, A.W. Wetzel, J. Gilbertson, M.J. Becich. Design and Analysis of Content-based Pathology Image Retrieval System. IEEE Transactions on Information Technology in Biomedicine, vol. 7, No.4, December 2003.