Background

Picture archive and communication systems (PACS) often contain millions of images related to clinical studies performed at an institution. In the current standard of practice, these images are often not annotated in a standards-based, structured fashion which would allow retrieval based on image features identified by an expert observer. The integration of content-based image retrieval (CBIR) into radiologist workstations is important for situations when relevant images cannot be retrieved by the coarse search of patient and study related information alone. A content-based medical image retrieval system could be used extensively for reference and teaching by providing retrieval of cases similar to a given image. This kind of system could also be used for decision support by retrieving cases with similar features and providing assisted diagnosis based on known pathology in retrieved cases. This paper describes a demonstration CBIR system for pulmonary nodule lookup.

Evaluation

The Lung Image Database Consortium (LIDC) has collected a database of lung CT images annotated by expert radiologists. Nodules in these images are outlined (regions of interest: ROI) and annotated with nine particular nodule features: calcification, internal structure, subtlety, lobulation, margin, sphericity, malignancy, texture and spiculation. All of these features are rated on an integer scale from 1 to 5 (except calcification, which is technically a subset of internal structure and is rated on a scale from 1 to 6).

The database is available online at https://imaging.nci.nih.gov/ncia/. At the time of our work there were 29 cases, each consisting of 100-400 DICOM images and an XML data file containing the annotations. We disregarded all nodules smaller than 5x5 pixels (around 3x3 mm) since objects of this size would not yield meaningful texture data. Our final collection consisted of 1106 annotations of 73 unique nodules. The median ROI size in pixels is 17x17 and the median actual size is approximately 11x11 mm. The smallest nodules are roughly 3x3 mm, while the largest are over 50x50 mm.

Since texture seems to be a primary characteristic used by radiologists to visually describe lung nodules, we were interested in how well texture feature analysis performs for content-based image retrieval on this dataset. We compared three different types of texture features: (1) Co-occurrence matrices, (2) Gabor filters, and (3) Markov random fields. These methods were used to extract a "feature vector" (a series of numbers) from ROIs that represent the nodule's signature. This vector was then compared with the vectors of other nodules by various similarity measures (Euclidean, Manhattan and Chebychev for point-based features; Chi-Squared and Jeffrey Divergence for histogram-based features).

We decided to base our evaluation 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, precision was defined as the number of retrieved instances of the query nodule divided by the number of retrieved images and recall was defined as the number of retrieved instances of the query nodule divided by the number 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.

Discussion

We have developed an application (see Figure 1) in C# which allows us to perform automatic lookup and similarity retrieval on the LIDC nodules. The application is fairly modular and this allows extension to include different descriptors and similarity measures. It provides for integration into existing workstation projects. The software is available as open source at http://brisc.sourceforge.net.

In our initial comparisons of system precision with respect to the similarity measure used, we found that the Euclidean similarity metric performed best for co-occurrence matrices while the Chi-Squared statistic

performed best for both Gabor and Markov calculations. Having determined the best similarity metric, we compared the descriptors using mean precision (average when performing a similarity query on every nodule in the database) with respect to the number of items retrieved, the size of the nodules, and radiologist agreement.

Figure 2 shows that when we vary the number of items retrieved, Gabor and Markov perform identically, with the best mean precision of 69% when one item is retrieved. Figure 2 also shows that Markov performs similarly to Gabor filters 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. This may be due to performing co-occurrence calculations at a global level while both Gabor and Markov are calculated at the pixel level. It is possible that pixel level co-occurrence matrix statistics will perform better.

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 3 shows the precision calculations for different nodule sizes. Figure 3 also shows that Markov and Gabor perform nearly identically and co-occurrence again performs worse. The graph shows that between the middle two ranges (approximately 7-11 mm and 11-16 mm) there is a 20 point difference in the precision for Gabor and a 7 point difference in the top two ranges (approximately 11-16 mm and 16-51 mm). This same trend is seen for the Markov statistics. These measures, therefore, work better on larger nodules.

Unfortunately, many of the physician ratings are inconsistent even when analysis is confined to a single nodule. Since this seems to imply that some nodules are hard to characterize even for trained radiologists, we ran precision calculations on nodules for which radiologists agreed on the "texture" annotation. As can be seen in Figure 4, when just two radiologists agree the mean precision jumped 21 points from 69% to 90% for both Gabor and Markov. Once three or four radiologists agree the precision is effectively 100%. Co-occurrence also performs significantly better when physicians agree, but the best precision is still only 55%.

Conclusion

We have developed a framework for content-based medical image lookup. When applied to the LIDC pulmonary nodule database, the system results in a mean precision of 69% (and higher when the query is limited to images on which physicians give consistent ratings). Given the high initial precision obtained by our algorithms, the system holds much potential for future refinement and integration into a radiologist workstation. Future improvements could include studies of the system's execution speed (since our dataset is relatively small, searches are nearly instantaneous) and the addition of data clustering algorithms for better indexing and more efficient retrieval. The system will be integrated into a workstation development project at Northwestern University.

Keywords

computed tomography, image feature, content-based image retrieval, pulmonary nodule

Figure 1: Nodule Viewer

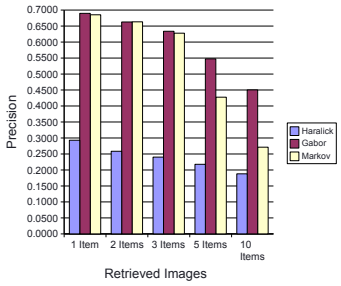

Figure 2: Images Retrieved

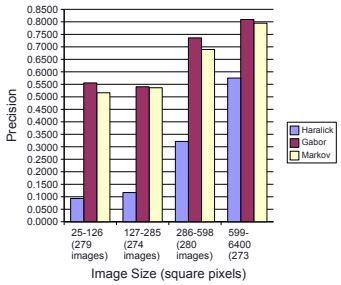

Figure 3: Image Sizes

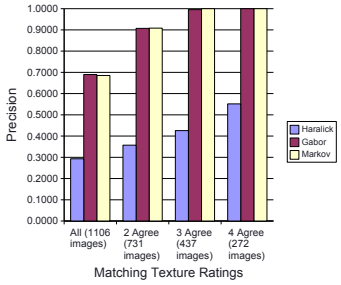

Figure 4: Physician Agreement