Content-based versus Semantic-based Retrieval: A LIDC Case Study

Sarah A. Jabon[a], Dr. Daniela S. Raicu[b], Dr. Jacob D. Furst[b]

[a]Rose-Hulman Institute of Technology, Terre Haute, IN, 47803
[b]School of Computing, CDM, DePaul University, Chicago, IL 60604

ABSTRACT

Content based image retrieval is an active area of medical imaging research. One use of content based image retrieval (CBIR) is presentation of known, reference images similar to an unknown case. These comparison images may reduce the radiologist's uncertainty in interpreting that case. It is, therefore, important to present radiologists with systems whose computed-similarity results correspond to human perceived-similarity. In our previous work, we developed an open-source CBIR system that inputs a computed tomography (CT) image of a lung nodule as a query and retrieves similar lung nodule images based on content-based image features. In this paper, we extend our previous work by studying the relationships between the two types of retrieval, content-based and semantic-based, with the final goal of integrating them into a system that will take advantage of both retrieval approaches. Our preliminary results on the Lung Image Database Consortium (LIDC) dataset using four types of image features, seven radiologists' rated semantic characteristics and two simple similarity measures show that a substantial number of nodules identified as similar based on image features are also identified as similar based on semantic characteristics. Furthermore, by integrating the two types of features, the similarity retrieval improves with respect to certain nodule characteristics.

Keywords: content-based image retrieval, CT scans, lung nodules, semantic-based image retrieval

I. INTRODUCTION

Lung cancer kills more people than any other cancer [1]. In 2008, the official estimate is that 215,020 cases will be diagnosed and 161,840 deaths will occur as a result of this disease. The five-year relative-survival rate between 1996 and 2004 was a mere 15.2% [3]. Early detection is critical to improving long-term survival and computed tomography is the premiere imaging modality for the detection of lung cancer, although screening with CT is still controversial.

Content-based image retrieval has the potential to improve diagnosis of malignant pulmonary nodules by providing the radiologist with images of similar nodules of known pathology. Many research groups have worked on CBIR for medical imaging. For instance, McNitt-Gray et al. used size, shape, and co-occurrence to classify a query nodule as malignant or benign using linear discriminant analysis [9]. Similarly, Armato et al. used nodule size and shape to also predict a query's malignancy [10]. Solely content-based systems can be very effective. However, it is difficult to accurately predict a human's interpretation of an image only with features extracted with computer algorithms. Hence, incorporating semantic information can have a substantial effect on the accuracy of a CBIR system.

Several studies incorporated this semantic information by making use of user feedback. Li et al investigated four different similarity techniques that include radiologists' ratings in a CBIR system for low-dose CT scans [6]. This system was developed using previously acquired similarity ratings by radiologists. Researchers at University of Illinois developed a CBIR system called BiasMap [8]. The neural-network-based BiasMap works with user feedback to improve a CBIR system. They developed their system based on discriminant analysis and it handles feedback data of small size, including uneven sizes of the positive and negative examples. However, more research is needed because currently, computer-aided diagnosis (CAD) systems that label the nodule as malignant or benign often have many false positives (3.8 false positives per case) [7].

In our previously developed CBIR system, BRISC [1], we ranked lung nodule similarity based on 64 content-based features. To ensure that a CBIR system has the potential to be used in clinical practice, it must be evaluated with respect to ground truth. Given that such a ground truth is not provided for the LIDC dataset (as the radiologists did not mark the similarity among the nodules), we evaluate BRISC in the context of the semantic characteristics used by radiologists to first describe a nodule appearance and then make the diagnosis. This kind of evaluation is more complex than the standard evaluation of CBIR systems when one radiologist looks at the nodules and marks them, for instance, as similar, non-similar, and neutral. In the context of the LIDC dataset [4], we generate the ground truth by quantifying the human perception of similarity based on calculated semantic-based similarities. In other words, if the values for the seven semantic characteristics of two nodules are close to each other, then those two nodules are considered to be semantically similar. Hence, in this paper we aim to evaluate the correlation between the content-based and semantic-based characteristics for the LIDC dataset and explore ways to improve the retrieval results based on correlations between the two.

II. METHODS

In this section we briefly explain the LIDC data set, the low-level image features and the semantic characteristics, and the similarity measures that we use to distinguish which nodules are most similar to others. Figure 1 presents the diagram of our methodology.

2.1 Data: Lung Images and Semantic Characteristics

The LIDC database contains complete thoracic CT scans for 85 patients along with XML files that contain the spatial locations of three types of lesions as marked by a panel of 4 LIDC radiologists. Any LIDC radiologist who identified a lesion as a nodule > 3 mm also provided subjective ratings for 9 nodule characteristics: subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture, and malignancy likelihood. For example, the texture characteristic provides information regarding the radiographic solidity of a nodule ("Non-Solid", "Part Solid/(Mixed)", "Solid") and the malignancy characteristic captures the likelihood of malignancy ("Highly Unlikely", "Moderately Unlikely", "Indeterminate", "Moderately Suspicious", "Highly Suspicious"). With the exception of the two characteristics that we excluded, calcification and internal structure, all semantic characteristics were ranked on a scale from one to five.

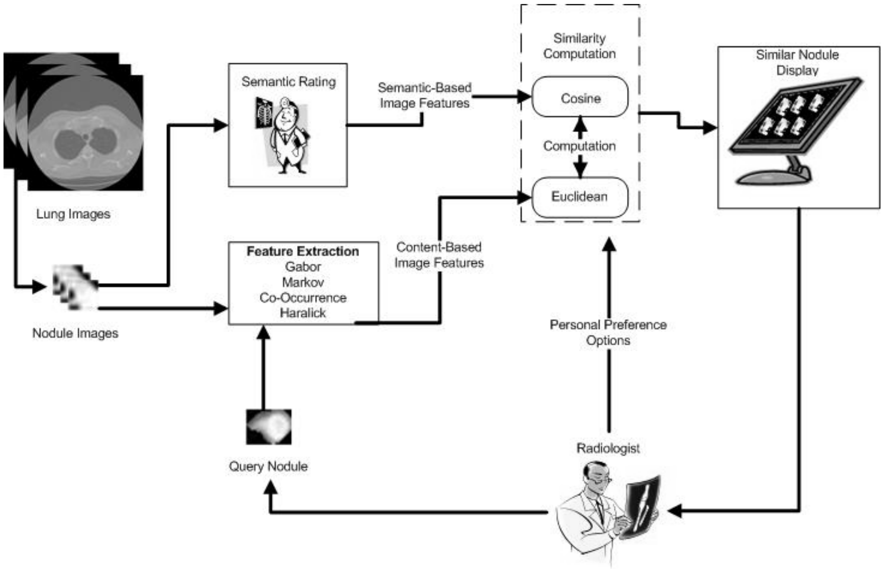
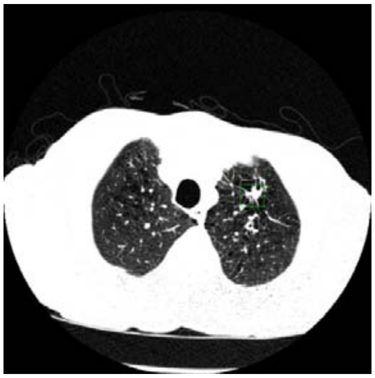


Figure 1: Diagram of the proposed methodology

While BRISC used all 1989 images (generated through the individual radiologist's outlines and the slices in which the nodules appear) for all lesions marked as nodules >3mm to assess the retrieval precision (number of retrieved images/instances of the same nodule divided by the total number of retrieved images in response to a query), in the proposed study we consider only one instance per nodule since radiologists are interested in retrieving images of other nodules and not from the same nodule. We reduce the dataset to 149 images by considering for each nodule the radiologist's outline in which the nodule has the largest area. The semantic characteristics ratings associated with that image – up to four values, depending on how many radiologists rated the image – are also aggregated to only one rating per characteristic and per nodule (Figure 2).



Rad.	Lob.	Mal.	Marg.	Spher.	Spic.	Subt.	Text.
A	3	4	4	2	4	3	4
B	4	3	4	4	3	5	5
C	4	2	3	4	3	4	5
D	4	3	2	2	4	3	3
Summarized	4	3	4	3	3	3	5

Figure 2: A CT image of the lung with a nodule specified in a box. Four radiologists' ratings for the nodule are in the table on the right, along with the final 7 semantic characteristics' ratings for that nodule.

Since the semantic ratings are ordinal categorical values, we chose the mode to encode the radiologists' ratings per characteristic. In the case that a unique mode did not exist, we used the median. Furthermore, if the median produced an overall rating that was not an integer number, then it was round down to the nearest whole number. Our new semantic-based vectors represented the radiologists' ratings relatively well. In fact, the average difference between our summarized vector and each radiologist's vector was only 2.618 out of a 28-point possible difference (a highest difference of 4 for 7 characteristics – calcification and internal structure were not included in the analysis because they had only one rating value for the entire dataset), which leaves us with a 9.4% error. Although the error may seem high, this is not an effect of the encoding technique for ratings summarization but rather a consequence of the high variability among radiologists when interpreting the data.

Figure 3 shows us that the lowest correlation coefficient was .578, which is a good result. The radiologists' themselves disagree quite a bit. In [11], though, we see that the average disagreement between raters themselves is .2 to .4. So, their agreement would be 1 minus the disagreement value: 0.6 to 0.8. The majority of the characteristics fall in that range when we take the mode or median according to our algorithm. Clearly, our summarization of the semantic-based characteristics is as accurate – if not more accurate – than another radiologist's rating.

Correlation Coefficients between the Summarized Vector and the Radiologists' Vectors	
Lobulation	0.5777
Malignancy	0.6481
Margin	0.7394
Sphericity	0.5945
Spiculation	0.6155
Subtlety	0.7567
Texture	0.7445

Figure 3: The correlation coefficients between the summarized feature vectors and the radiologists' individual vectors

2.2 Low-level Image Features

For each image, we calculated 64 different content-based features: 8 shape features (circularity, roughness, elongation, compactness, eccentricity, solidity, extent, and standard deviation of radial distance); 7 size features (area, convex area, perimeter, convex perimeter, equivalence diameter, major axis length, and minor axis length), 5 gray-level intensity features (minimum, maximum, mean, standard deviation, and difference), and 44 texture features (based on co-occurrence matrices, Gabor filters, and Markov random fields). For full explanations of all these concepts, please see [1]. We now can define the query nodule (Q) and the database nodule (N) respectively:

$$\mathrm{Q} = (\mathrm{c}_1^{\mathrm{Q}}, \mathrm{c}_2^{\mathrm{Q}}, \dots \mathrm{c}_7^{\mathrm{Q}}, \mathrm{f}_1^{\mathrm{Q}}, \mathrm{f}_2^{\mathrm{Q}}, \dots \mathrm{f}_{64}^{\mathrm{Q}})$$
$$\mathrm{N} = (\mathrm{c}_1^{\mathrm{N}}, \mathrm{c}_2^{\mathrm{N}}, \dots \mathrm{c}_7^{\mathrm{N}}, \mathrm{f}_1^{\mathrm{N}}, \mathrm{f}_2^{\mathrm{N}}, \dots \mathrm{f}_{64}^{\mathrm{N}})$$

where c is a semantic-based characteristic as rated by radiologists and *f* is a content-based feature assessed by the algorithms used in BRISC [1].

2.3 Similarity Measures

For the semantic-based similarity, we considered several similarity measures, such as the extended Jaccard (Tanimoto coefficient) and the cosine measure. We used the cosine similarity measure because it minimized the ceiling effect in our data; with other measures, the majority of the nodules were labeled as similar to the query nodule – within 0.0001 points in a 0 to 1 scale. Although there is still a ceiling effect with the cosine similarity, it is substantially minimized. In order to make the least similar nodules have a value of 1 and the most similar nodules have a value of 0, we subtracted the cosine measure from 1. The final formula is shown below.

$$S_c(Q,N) = 1 - \frac{\sum_{i=1}^{7} c_i^Q * c_i^N}{\sqrt{\sum_{i=1}^{7} (c_i^Q)^2} * \sqrt{\sum_{i=1}^{7} (c_i^N)^2}} \quad (1)$$

For the content-based similarity, we considered the Euclidean distance based on our previous results [1]. The Euclidian distance is described below:

$$S_E(Q,N) = \sqrt{\sum_{i=1}^{64} (f_i^Q - f_i^N)^2} \quad (2)$$

We first computed the similarity between each vector using the Gabor features, the Markov features, and the co-occurrence features individually. Then, we computed the similarity measures for those three features combined. Finally, we used all the features we have for each image. We normalized all the content-based similarity values using the min-max normalization technique [5].

We also calculated a similarity based on the weights of each feature. This weighted Euclidean formula is below:

$$S_W(Q,N) = \sqrt{\sum_{i=1}^{64} w_i * (f_i^Q - f_i^N)^2} \quad (3)$$

The weight for each of the 64 features was the absolute value of the correlation coefficient for the corresponding semantic characteristic. We only used this similarity measure when working with individual semantic features and correlations (Section 3.3).

III. RESULTS

In order to assess the correlation between the two similarity measures, we used a round robin approach where we extracted one nodule as a query and compared it to the remaining 148 nodules. We took the *k* most similar values from each query's semantic-based similarity ordered list and content-based similarity ordered list and counted how many nodules were common to both lists.

3.1 Using a Set Number of Nodules to Determine the Number of Matches

In our first method, we used a set number of nodules in the list. Using *k*=20 for the number of most similar images to be compared between the two approaches, we found the combination of the three texture models and the combination of all 64 low-level features resulted in a higher number of matches than the individual features (Figure 4). Both had a substantially more nodules with 6-10 matches in the set of 20. The number of nodules that had 2 - 5 matches was relatively consistent throughout all image features, but slightly higher for Gabor and Markov. No combination of image features had more than 10 matches out of the twenty most similar. Based on these results, for the rest of the experiments, we will work with all 64 image features.

Matches	Gabor	Markov	Co-Occurrence	Gabor, Markov, and Co-Occurrence	All Features
6 – 10	24	18	31	36	43
2 – 5	107	104	94	98	93
0 – 1	18	27	24	15	13

Figure 4: Match Counts in 20 Most Similar Nodules

In order to analyze the results with respect to the retrieval using a similarity threshold, we examined the histograms of the semantic-based similarity values, content-based similarity values, and the correlations between the two (Figure 5).

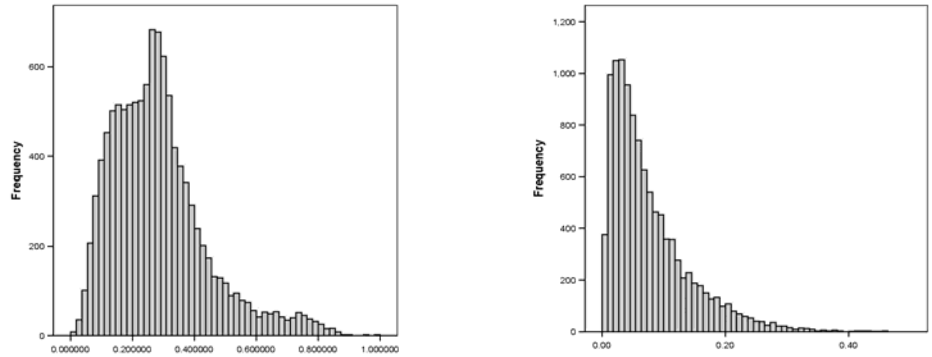


Figure 5: Left: Histogram of content-based similarities; right: Histogram of semantic-based similarities

We notice that although the possible range for the semantic-based similarity is 0 to 1, all the values are in the range 0 to 0.4 with a bell curve distribution skewed to the right. Given the fact that 0 denotes perfect similarity, we can say that most of the nodules are similar with each other based on the semantic characteristics and, if there is any dissimilarity, it might be difficult to perceive. Looking at the content-based histogram, we notice almost a normal distribution with the center around 0.37 and a longer tie to the right indicating nodule pairs dissimilar to each other.

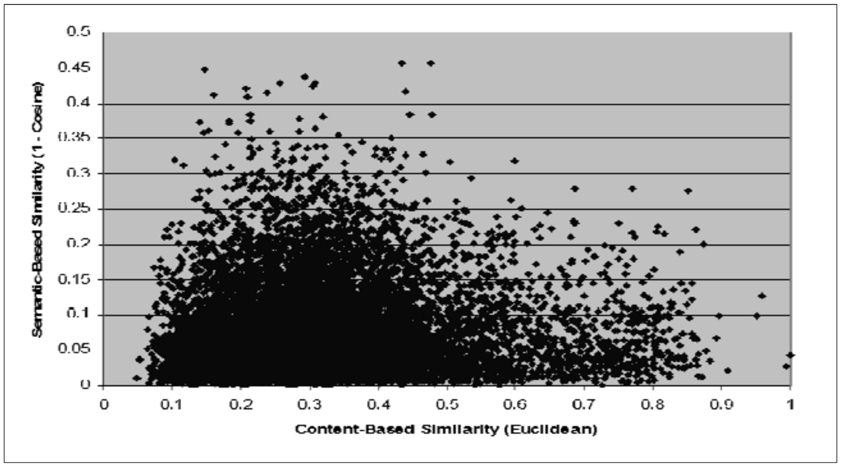


Figure 6: Scatter plot of the 11,026 pair-wise content-based and semantic based values

In order to compare the nodule corresponding similarity values, a scatter plot is useful (Figure 6). Analyzing the scatter plot, we can see that there is a correspondence between the content-based and semantic-based similarity values in terms of their monotony. We expect the semantic-based similarity to be similar to the content-based similarity, which would result in a trend line that has a slope of one. However, the similarity based on the semantic-based features has a substantial ceiling effect which is difficult to predict with the computer extracted features.

3.2 Using a Threshold to Determine the Number of Matches

Based on these observations, we investigated the number of matches with respect to a certain threshold instead of choosing to only compare within the 20 most similar nodules. We also analyzed the box plots of the similarity values of the matches out of the top twenty to choose what thresholds to use; we decided to use a threshold of 0.2 for the content-based similarity because it included 75% of the nodules for which there were 7, 8, or 9 matches. For the semantic-based characteristics, we tried a few different thresholds: 0.04, 0.02, and 0.001. The threshold of 0.02 had up to 31 nodule matches out of all 149, where the threshold of 0.04 had up to 56 matches (Figure 7).

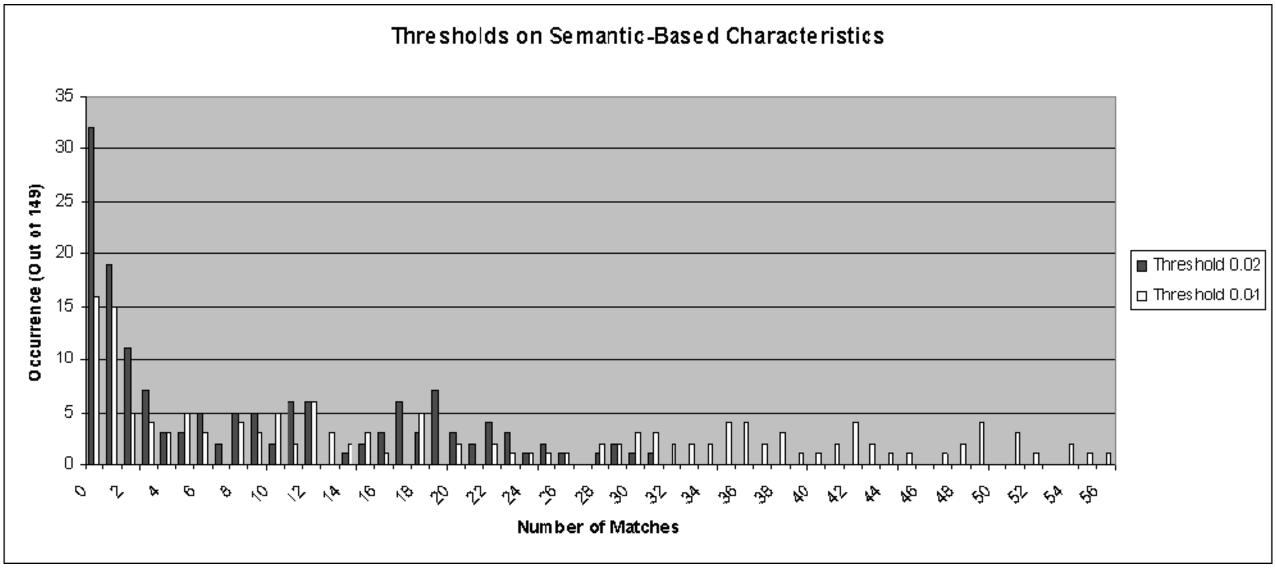


Figure 7: Matches based all features and using thresholds

The nodules that have the most matches have characteristics in the middle of the 1 – 5 range. The nodule with 31 matched has the ratings four 3’s, two 4’s, and one 2 for the seven ratings. It makes sense that such a nodule would return many similar nodules, as the extremes of the ratings on the characteristics are not far off whether they are high or low. The nodules with the most matches for the 0.02 threshold and the 0.04 threshold are show below (Figure 8).

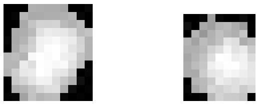


Figure 8: The nodule with 31 matches with a 0.02 threshold (left) and the nodule with 56 matches with a 0.04 threshold (right).

The threshold of 0.001 produced very low results – 142 nodules with zero matches, 5 nodules with one match, and 2 with two matches. These results show that by not restricting the retrieval results to a certain number of retrieved nodules, the number of matches between the two approaches can be substantially large depending on the threshold used on the semantic-based similarities.

3.3 Weighting Individual Semantic Characteristics to Determine the Number of Matches

We also incorporated the semantic information in the content-based approach. We weighted each one of image features by their correlation with a semantic characteristic of interest. For instance, if lobulation was 5, all the nodules that had a rating of 5 were most similar, those with a rating of 4 were second-most similar, etcetera. Then, within the nodules with a specific rating, we rated their similarity by the cosine similarity for all seven semantic-based characteristics. This created a list that was filtered by a specific characteristic, but still took into account the similarity for all semantic-based characteristics. The results for the content-based approach were then compared with the ones for the semantic-based approach filtered by the characteristics of interest.

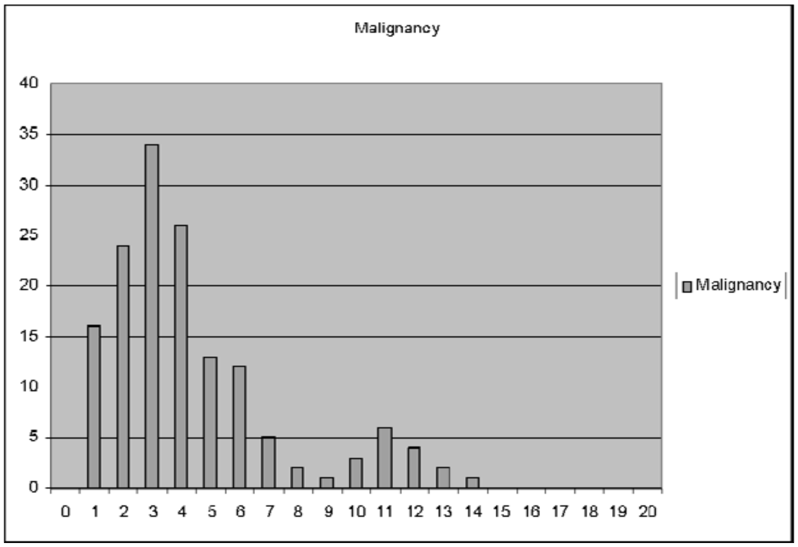


Figure 9: Matches based on weighting malignancy

Malignancy had up to 14 matches out of 20 when using the correlations as weights (Figure 9). There were only 16 nodules that had one match, and 36 nodules that had more than five matches. Although 36 is not as high as the number of nodules that have 6 - 9 matches with all seven characteristics, filtering with malignancy results in a higher number of matches for numerous nodules.

Using this simple way of integrating the two approaches, the number of matches for malignancy and subtlety was even 14 out of 20 matches for seven of the nodules. This may be explained with the high correlation between these features and the content-based features. The average absolute value of the correlation is 0.251 and 0.207 for malignancy and subtlety respectively, whereas the average correlations of the other characteristics were between 0.11 and 0.58. We plan to explore this scheme further for content and semantic-based features integration.

IV. CONCLUSIONS

Currently, content-based image retrieval is solely based on image features identified by computer-based algorithms. These features may or may not represent what radiologists want to see when determining the similarity between nodules. In order to ensure that these features are representative of the similarity that the radiologists use to characterize nodules, it is essential to correlate the content-based similarity well with the semantic-based similarity. In this paper, we present the evaluation of our CBIR for the LIDC dataset in the context of the semantic characteristics used by radiologists to first describe a nodule appearance and then make the diagnosis. This kind of evaluation is more complex than the standard evaluation of CBIR systems when ground truth from one radiologist is available for evaluation.

Our preliminary results show that a substantial number of nodules identified as similar based on image features are also identified as similar based on semantic characteristics and therefore, the image features capture properties that radiologists look at when interpreting lung nodules. Further research is necessary to investigate further the correlations between the two types of features and integrate them in one retrieval system that will be of clinical use.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0755407.

REFERENCES

[1] Lam, M., Disney, T., Pham, M., Raicu, D., Furst, J., "Content-Based Image Retrieval for Pulmonary Computed Tomography Nodule Images", SPIE Medical Imaging Conference, San Diego, CA, February 2007.

[2] Zhou, X., Huang, T., "Relevance Feedback in Image Retrieval: A Comprehensive Review", IEEE CVPR2001 Workshop, 2001.

[3] Ries, LAG, Melbert, D., Krapcho, M., Stinchcomb, D.G., Howlader, N., Horner, M.J., Mariotto, A., Miller, B.A., Feuer, E.J., Altekruse, S.F., Lewis, D.R., Clegg, L., Eisner, M.P., Reichman, M., Edwards, B.K. (editors). SEER Cancer Statistics Review, 1975-2005, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2005/, based on November 2007 SEER data submission, posted to the SEER web site, 2008.

[4] The National Cancer Institute, "Lung Imaging Database Consortium (LIDC), http://imaging.cancer.gov/programsandresources/InformationSystems/LIDC.

[5] Han, J., Kamber, M., [Data Mining: Concepts and Techniques], London: Academic P, 2001.

[6] Li, Q., Li, F., Shiraishi, J., Katsuragwa, 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", *Medical Physics* 30:2584-2593, 2003.

[7] Doi, K., "Current Status and Future Potential of Computer-Aided Diagnosis in Medical Imaging," The British Journal of Radiology, 2005.

[8] Zhou, X.S., Huang, T.S., "Small sample learning during multimedia retrieval using BiasMap," IEEE Conf., Computer Vision and Pattern Recognition, Hawaii, 2001.

[9] McNitt-Gray, M.F., Hart, E.M., Wyckoff, N., Sayre, J.W., Goldin, J.G., Aberle, D.R., "A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: Preliminary results," Medical Physics, vol. 26, no. 6, pp. 880–888, 1999.

[10] Armato, S.G. III, Altman, M.B., Wilkie, J., Sone, S., Li, F., Doi, K., Roy, A.S., "Automated lung nodule classification following automated nodule detection on CT: A serial approach." *Medical Physics* 30: 1188–1197, 2003.

[11] Horsthemke, William H., D. S. Raicu, J. D. Furst, "Bridging the Evaluation Gap Challenge to Diagnostic Labeling of Pulmonary Nodules", DePaul CDM Research Symposium, 2008.