An Investigation into the Relationship between Semantic and Content Based Similarity using LIDC

Robert Kim
Johns Hopkins University
Baltimore, MD 21218
robert.kim@jhu.edu

Grace Dasovich
Northwestern University
Evanston, IL 60201

Runa Bhaumik
DePaul University
School of Computing, CDM
Chicago, IL 60604

Richard Brock
DePaul University
School of Computing, CDM
Chicago, IL 60604

Jacob D. Furst
DePaul University
School of Computing, CDM
Chicago, IL 60604
jfurst@cdm.depaul.edu

Daniela S. Raicu
DePaul University
School of Computing, CDM
Chicago, IL 60604
draicu@cdm.depaul.edu

ABSTRACT

There is considerable research in the field of content-based medical image retrieval; however, few of the current systems investigate the relationship between the radiologists' visual impression of image similarity and the computer calculated content-based similarity. Furthermore, those research studies that investigate these relationships analyze the visual similarity with respect to degree of malignancy without including specific characteristics that are important in the diagnosis process.

The creation of the NIH/NCI Lung Image Database Consortium (LIDC) dataset offers the opportunity to perform the proposed research. Each nodule out of the 932 distinct nodules (larger than 3mm in diameter) was delineated and annotated by up to four radiologists using nine semantic characteristics that are important in the lung nodule interpretation process. Using the LIDC images, we propose to encode the radiologists' characteristic-based similarity and further discover if there is any relationship between this conceptual/characteristic-based similarity and the content-based similarity for lung nodule interpretation.

Our preliminary results show that it is a challenging problem to model the characteristic-based and content-based relationships for a broad category of lung nodules. A correlation of only 0.1 was obtained between the characteristic-based similarity and the predicted characteristic-based similarity using an artificial neural networked trained on four types of low-level image features (size, intensity, shape, and texture) calculated for 640 random pairs of nodules. Future research is necessary to investigate the appropriateness of the considered image features to model both the variation in the human interpretation of the lung nodules and the perceived characteristic-based similarity.

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.
Copyright 2010 ACM 978-1-60558-815-5/10/03 ...$10.00.

Categories and Subject Descriptors

J.6 [**Computer-Aided Engineering**]: Computer-aided design (CAD)

General Terms

Performance

Keywords

CT scans, Content based image retrieval, lung nodules, semantic based image retrieval

1. INTRODUCTION

Lung cancer has one of the shortest survival rates after diagnosis of all cancers[12]. Studies show that the presence of similar images assists radiologists in correctly diagnosing lung nodules as benign or malignant[13]. Content-based image retrieval (CBIR) systems, where image features of a query image are compared to a database of images, are employed to obtain similar images. One such system is BRISC, implemented in previous work by Lam et al.[11] Lam computed 40 image features using three texture models: Haralick co-occurrence matrices, Gabor filters, and Markov random fields to construct a CBIR system for pulmonary nodules. Furthermore, Lam found that image features generated from Gabor texture model produced the best retrieval results.

However, content-based similarity obtained from computer algorithms does not necessarily correspond to a human perception of image similarity. In order to bridge this gap, Jabon et al.[9] expanded the BRISC project by taking into account both content and semantic based features. Using the National Institutes of Health (NIH) Lung Image Database Consortium (LIDC) images, Jabon compared the content-based retrieval using 64 image features and the semantic based retrieval using four radiologists' ratings on seven nodule characteristics. Jabon discovered that a substantial number of nodules recognized as similar semantically were also similar based on image features.

In this paper, we extend work by Jabon by constructing a model using an artificial neural network (ANN) between the two types of retrieval systems, content-based and semantic-based, for several subsets of LIDC lung nodule pairs.

First, we encode the radiologists' characteristic-based similarity by using and evaluating two probabilistic-based similarity measures: Jeffrey divergence and the Earth Mover's distance. Our proposed probabilistic-based similarity approach allows taking into account the variability of radiologists' when assessing the degree of likelihood for each characteristic (such as 1="extremely subtle" to 5="obvious" for subtlety characteristic).

Second, we encode the content-based similarity measures using the most used low-level image features for lung nodule interpretation that were found to be important when classifying nodules with respect to malignancy and the other semantic characteristics based on our previous research work. The absolute difference of individual features is used to encode the content-based similarity of nodules.

Third, the relationships between these two types of similarities, semantic based and content based, are investigated using a neural network approach in which the characteristic-based similarity is the teaching signal and the content-based similarity is the input signal.

A successful neural network prediction model can be utilized to predict the semantic similarity for new nodules that have not been annotated by radiologists. When new nodules are discovered, their low-level image features (content-based similarity) can be calculated, and these features can be used in conjunction with the neural network (that has been trained with the LIDC nodules with annotations) to derive a similarity close to the human perceived similarity quantified through the LIDC characteristics.

2. RELATED WORK

Numerous Computer-Aided Diagnosis systems have been developed in recent years for detection (CADd) and diagnosis (CADx) of pulmonary nodules and interstitial lung disease in chest radiography and CT. Several researchers [3, 4, 7] showed that artificial neural networks can provide powerful tools in the diagnosis of interstitial lung diseases. Other work [16, 22, 8] showed that texture features can be used to detect interstitial lung diseases. In the realms of lung nodule detection and diagnosis, the focus of our research, there are also several research efforts that show the promises of the CADd and CADx as 'second readers' in the lung nodule interpretation and decision making process. One such system was built by Armato et al.[2] who set up an automated classification based on a linear discriminant analysis (LDA) to differentiate malignant and benign lung nodules in low-dose computed tomography (CT). In their study, the features shape characteristics of lung nodules were merged through a linear discriminant classifier to classify the nodules. Although these studies illustrate that low-level image features can be used to distinguish between malignant and benign nodules, it is important to incorporate radiologists' knowledge into the process and to understand the relationship between the image features and radiologists' annotations. Such understanding can not only improve diagnosis of malignant lung nodules, but also simplify and accelerate the radiology interpretation process as suggested by Kahn et al.[10]

In the medical imaging area, efforts to find the relationship between image features and subjective or semantic ratings were spearheaded by Barb et al.[5], Raicu et al.[19], and Samala et al.[21] . Barb developed a framework that manages visual content of lung pathologies. The framework named Evolutionary System for Semantic Exchange of Information in Collaborative Environments (ESSENCE) uses semantic methods to describe visual abnormalities and exchange knowledge in the medical domain. The framework largely consists of a semantic domain, a feature domain, and a preference domain. The semantic domain contains semantic ratings assigned by users (physicians), a feature domain stores image features extracted by feature extraction algorithms, and the preference domain contains user preferences. Fuzzy logic techniques were employed to map from low-level image features to high-level semantic terms and to retrieve images using both computed image features and physician-defined semantics. More recently, Raicu et al. developed two semi-supervised methods to predict semantic ratings for lung nodules from the Lung Image Database Consortium (LIDC) given low-level image features. Using an ensemble of classifiers from DECORATE[15] and decision trees, they were able to improve the accuracy prediction of semantic ratings by 50 % on average despite the variability in the radiologists' interpretation. Samala used nonparametric correlation coefficients, multiple regression analysis, principal-component analysis, and artificial neural network analysis to investigate the optimum selection of image features. They used 42 cases (28 lung nodules and 14 non-nodules) from LIDC for the feature characterization, and a total of 11 features were computed. Correlation analysis and multiple regression analysis were used to find the relationship between radiologists' ratings and the computed features, and a three-layer feed-forward neural network was used to classify the abnormal and normal lung nodules. The correlation coefficients between the radiologists' annotations on 9 characteristics and 11 computed image features range from 0.2693 to 0.5178. Using the multiple regression analysis, Samala et al. discovered that the higher number of features, the higher squared multiple correlation coefficients (R^2).

While the above work focuses on the relationship between the image features and the semantic characteristics in terms of their prediction power (for example how well image features can be used to predict the spiculation perception of a lung nodule in Raicu et at.[19]), there is some work that focuses on finding the same relationships but with respect to how closely the image features capture the human perception of similarity. For mammographic masses, much work has been done to establish the relationship between image features and radiologists' similarity perception. Muramatsu et al.[17] used an ANN to find the relationship between the image features and the radiologists' ratings on mammograms. The subjective similarity ratings for 300 pairs of images with clustered microcalcifications were obtained from ten radiologists and the average values of these ratings were used as teaching data for a three-layer feed-forward ANN with a backpropagation algorithm. Seven image features were used as inputs and the ANN was trained to predict the semantic similarity called a psychophysical similarity measure. The correlation coefficient between the radiologists' ratings and the psychophysical similarity measure was 0.71.

More recently, preliminary work by Muramatsu et al.[18] links image features to the Breast Imaging Reporting and Data System (BI-RADS). Using an artificial neural network (ANN), Muramatsu et al. determined similarity measures between subjective features (BI-RADS descriptors assigned by radiologists) and objective features (computed image features) for pairs of breast masses. The ANN was trained with average ratings by 10 breast radiologists as teaching data and the BI-RADS lesion descriptors or image features as input data. Several feature combinations were tested, and the leave-one-out method was used to test the ANN. Muramatsu et al. found that when the BI-RADS descriptors were used as input data for the ANN, the correlation coefficient was decent. However, when the combinations of image features and the BI-RADS descriptors were used as input, the correlation coefficients were relatively high.

For pulmonary nodules, there is not much work done to investigate the correlation between the computer similarity results and the radiologists' similarity perception. Li et al.[13] computed similarity measures using four different techniques: feature-based, pixel-value-difference-based, cross correlation based, and ANN-based techniques. They discovered that the Artificial Neural Network (ANN) technique gave the highest correlation, 0.72. The input layer takes seven content-based image features, while the semantic similarity ratings by ten radiologists were used as a teaching signal. 240 pairs of lung nodules were used for the ANN, and the leave-one-out method was applied to verify the effectiveness of the ANN. The proposed approach is most similar to Li's work, however, while Li investigates the absolute similarity for nodule images, we propose to investigate a relative similarity with respect to semantic concepts that are used by LIDC radiologists to interpret lung nodules in the process of diagnosis.

Recently we developed a multivariate linear regression model using nodules pairs from the previously available (before June 2009) LIDC that contained the first 149 nodules from the latest LIDC data used in this paper[6]. The Cosine Similarity measure and the Euclidean distance were used to encode the semantic-based similarity and the content-based similarity, respectively. After evaluating these two types of similarities for all nodule pairs (11026 nodule pairs from the 149 distinct nodules), we selected 116 nodule pairs with high correlation between both similarities. In turn, these pairs were used to generate a linear regression model that predicts semantic similarity with content similarity input with an R^2 value of 0.871. In that work, we assumed a linear relationship between the semantic-based and the content-based similarities, and we did not take account of the radiologists' variability (Cosine similarity measure is not probabilistic-based). In this paper, we investigate a non-linear model using a neural network and probabilistic-based similarity measures, Jeffrey Divergence and Earth Mover's Distance methods.

3. MATERIALS AND METHODS

3.1 Semantic Ratings and Low-Level Image Features

The data used in this research was provided by the Lung Imaging Database Consortium (LIDC). The latest database, which was released in 2009, includes 399 unique sets of Computerized Axial Tomography (CT) scans of the lungs. Each set of scans was analyzed by up to four expert radiologists, and any nodules found were delineated and rated based on 9 semantic characteristics (calcification, internal structure, lobulation, malignancy, margin, sphericity, spiculation, subtlety, and texture)[1]. Out of the 399 sets of scans, 932 distinct nodules were found and analyzed. Most of these nodules appeared on multiple slices throughout a set of scans. In order to reduce the number of instances of each nodule to one, we calculated the area of the nodule on each slice using the radiologists' boundaries, and used the slice that contained the largest area of the nodule. If multiple radiologists delineated the nodule, the radiologist-outlined boundary including the maximum number of pixels was used to calculate the area of the nodule region.

Only seven of the characteristics were used because two of them, calcification and internal structure, had very little variation. Previous work extracted 64 image features based on four categories: texture (Gabor, Markov Random Fields, and Haralick Co-Occurrence), size, shape, and intensity[11] shown in Table 1. Each feature for the 932 nodules was normalized using the Z-Score method.

3.2 Similarity Measures

To determine the semantic-based similarity using all radiologists' ratings, the Jeffrey divergence and Earth Mover's Distance were used. They were selected because both methods incorporate multiple radiologists' ratings of a nodule.

The Jeffrey divergence measures the extent to which two probability distributions agree[14]:

$$S_J(m,n) = \sum_{i=1}^{7}\sum_{j=1}^{5}\left(m_{ij}\left|\log\left(\frac{m_{ij}}{\hat{P}_{ij}}\right)\right| + n_{ij}\left|\log\left(\frac{n_{ij}}{\hat{P}_{ij}}\right)\right|\right) \tag{1}$$

$$\text{where } \hat{P}_{ij} = \frac{m_{ij}+n_{ij}}{2}$$

$S_J(m,n)$ is the Jeffrey divergence between two nodules m and n using seven semantic features $i = 1, 2, \ldots, 7$ and five ratings ($j = 1, 2, \ldots, 5$) for each semantic feature. In the equation, m_{ij} or n_{ij} represents a probability distribution for feature i and rating j for nodule m or n. Because of the nature of the equation, nodules with ratings from only one radiologist have the same values of the Jeffrey divergence when compared with other nodules having only one set of ratings. Because of this, all nodules with only one set of ratings were removed for the neural network: 330 nodules were removed from the 2009 data set leaving 602 nodules. Figure 1 shows histograms of pairs from the LIDC data set that contains the first 149 nodules before and after the nodules with one set of ratings are removed (the previous LIDC data, known as LIDC85, contains the first 149 nodules of the 2009 data set).

Earth Mover's Distance is a method of comparing the distance between two distributions. It is computed by calculating the minimum cost of transforming the histogram of one element into the histogram of another. Earth Mover's Distance also relies on a separate "ground distance" function for which we used Jeffrey Divergence. A discussion of the Earth Mover's Distance as a metric for image retrieval can be found in [20].

$$S_{\text{emd}}(m,n) = \sum_{i=1}^{7}\sum_{j=1}^{7} d_{ij} f_{ij} \tag{2}$$

$S_{\text{emd}}(m,n)$ is the Earth Mover's Distance between nodule m and nodule n. In this distance measure, each nodule is represented by one signature, which consists of the probability distribution for each of the seven semantic features.

Table 1: All 64 computed image features

Shape Features	Size Features	Intensity Features	Texture Features
Circularity	Area	MinIntensity	11 Haralick features calculated from
Roughness	ConvexArea	MaxIntensity	co-occurrence matrices (Contrast,
Elongation	Perimeter	MeanIntensity	Correlation, Entropy, Energy,
Compactness	ConvexPerimeter	SDIntensity	Homogeneity, 3^{rd} Order Moment,
Eccentricity	EquivDiameter	MinIntensityBG	Inverse Differential Moment, Variance, Sum Average,
Solidity	MajorAxisLength	MaxIntnsityBG	Cluster Tendency, Maximum Probability
Extent	MinorAxisLength	MeanIntensityBG	24 Gabor features (mean and standard deviation of
RadialDistanceSD		SDIntensityBG	Gabor filters of four orientations and three scales)
		IntensityDifference	5 Markov features

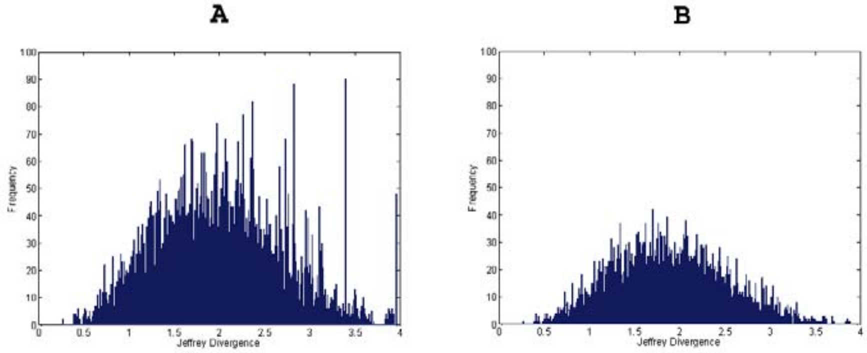

Figure 1: Histograms of the Jeffrey divergence for all 11,026 pairs from 149 distinct nodules (A) and the 5,886 pairs from 109 distinct nodules without nodules of only one set of ratings (B).

d_{ij} is the ground distance function between m_i and n_j. f is the flow that minimizes the overall cost, which is found by solving an instance of the Transportation problem.

While for concept based retrieval Jeffrey divergence and Earth Mover's Distance are investigated, for the content-based retrieval system the absolute difference between the features of nodule i and j, $d_n(i, j)$, was used:

$$d_n(i, j) = |f_n^i - f_n^j| \tag{3}$$

3.3 ANN Prediction Model

A prediction model using an artificial neural network was constructed to predict the semantic similarity from the computed image features for various selections of images.

For this model, three different subsets of pairs were used: pairs selected randomly, pairs from the largest 25% of nodules, and pairs from nodules rated moderatly or highly suspicious for malignancy by radiologists.

Instead of working with all 867,692 pairs, to reduce the computational time we selected 640 random pairs of nodules in the three different ways outlined above. We decided to work with 640 pairs for this preliminary study to avoid the curse of dimensionality problem, and therefore allow 10 different cases for each one of the 64 image features. The first subset was made up of 640 pairs randomly selected. For the largest 25% of nodules, there were total 184 unique nodules (out of 602 distinct nodules) whose area was larger than 93 mm^2 forming 16,836 pairs. Out of these 16,836 pairs we selected again randomly 640 pairs for the second subset. For the third subset, suspicious nodules were found by first taking the mode of the radiologists' ratings for malignancy and selecting nodules whose mode was greater than 3 (See Appendix A for definitions of the radiologists' ratings). This method gave 97 distinct nodules forming 4,656 pairs. Again random 640 pairs from the 4,656 pairs were selected to save computation time and used to train the neural network.

A three-layer, feed-forward neural network with a back-propagation algorithm was employed to learn the relationship between the computed image features and the radiologists' ratings. The differences in feature values (all 64 features) for a pair of nodules were used as input data. The single output represents a predicted semantic similarity for the pair. In the hidden layer, five neurons were used. Figure 2 illustrates our neural network.

During the training of the ANN, either the Jeffrey divergence similarity values or the Earth Mover's Distances were used as teaching data for the network. The ANN was trained with a hyperbolic tangent transfer function up to 200 iterations. To determine the predicted semantic similarity measures for all the pairs in a subset, a leave-one-out method was used. For example, consider the 640 pairs chosen randomly. One pair is extracted from the 640 pairs and the ANN is trained with the remaining 639 pairs. At the end of the training, the differences in the features for the excluded pair are used as inputs for the ANN which will return an output that represents the predicted semantic similarity. This process is repeated for each of the 640 pairs. The performance of the ANN was evaluated in terms of Pearson's correlation coefficient between the predicted semantic similarity values from the ANN and the actual semantic similarity values. The higher the correlation coefficient, the better is the accuracy of the ANN in predicting the semantic similarity.

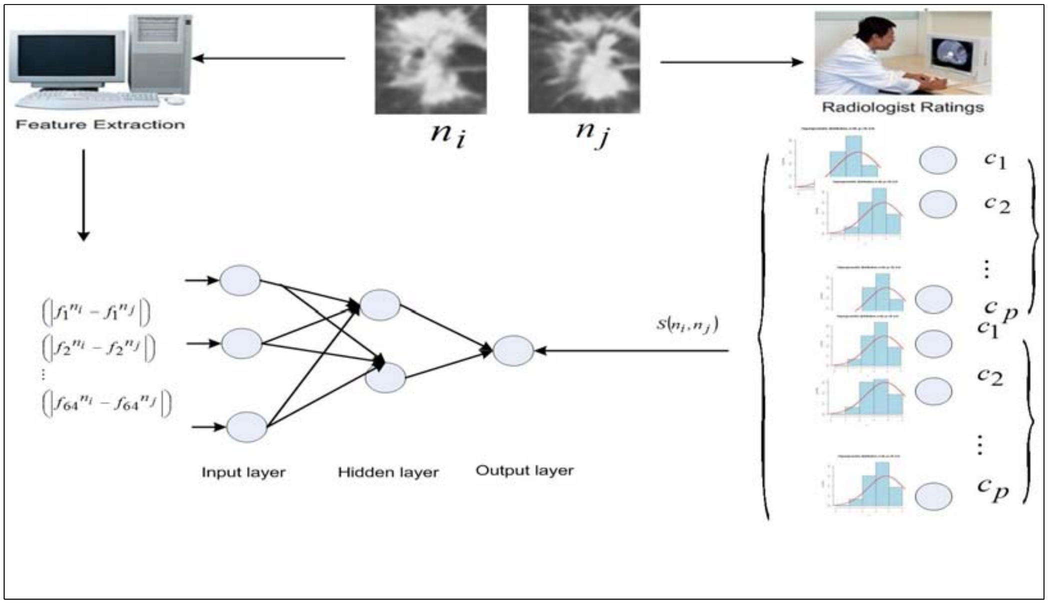

Figure 2: The diagram of the system including the topology of the neural network

Table 2: Correlation results between human perceived and predicted characteristic based similarity

	Semantic Distance	Content Distance	Correlation
640 Random Pairs	Jeffrey	Absolute difference	0.0644
Large Pairs	Jeffrey	Absolute difference	0.0387
Malignant Pairs	Jeffrey	Absolute difference	0.188
640 Random Pairs	Earth Mover's Distance	Absolute difference	0.129
Large Pairs	Earth Mover's Distance	Absolute difference	0.0385
Malignant Pairs	Earth Mover's Distance	Absolute difference	0.128

4. RESULTS

The correlation between the Jeffrey divergence values and the Euclidean distances for all 180901 pairs was 0.223. The correlation between the semantic similarity based on Jeffrey divergence and the content similarity using Euclidean distance for the random 640 pairs was very low, 0.0385. Using the ANN prediction model, we were able to improve the correlation between semantic and content similarity to 0.0644 which is still extremely low. The correlation between the Jeffrey divergence values and the Euclidean values for the highly malignant pairs was 0.0617, and for the largest 25% by area nodules 0.00780. The correlations between actual and predicted semantic similarity were increased to $r = 0.188$ and $r = 0.0387$, respectively.

When the Earth Mover's Distance was used to encode the semantic ratings, the correlations did not change much. For the random 640 pairs, the correlation went up to 0.129. For the largest 25% by area nodules, the correlation stayed about the same, 0.0385, while the correlation for the highly malignant pairs dropped to 0.128. The results are shown in Table 2.

Figure 3 shows examples of both semantic and content based image retrieval using an LIDC DICOM Analyzer created by Rick Brock.

5. CONCLUSION

Using an artificial neural network we predicted semantic similarity using similarity based only on computed features for a random selection of nodule pairs with the highest correlation of 0.129. Using the large or malignant pairs did not improve the correlation.

The neural network was expected to perform nearly as well as the linear model for all nodule pairs. The low correlation indicates the semantic gap still remains for nodule similarity. It may be that the image features thought to accurately represent a radiologists' interpretation of nodule characteristics do not correspond to similarity. Research indicates that multiple features work better in combination, however too many features might results in over fitting. Thus far in terms of similarity the right combination is unknown.

This research used only traditional methods of machine learning. In the future alternate methods should be inves-

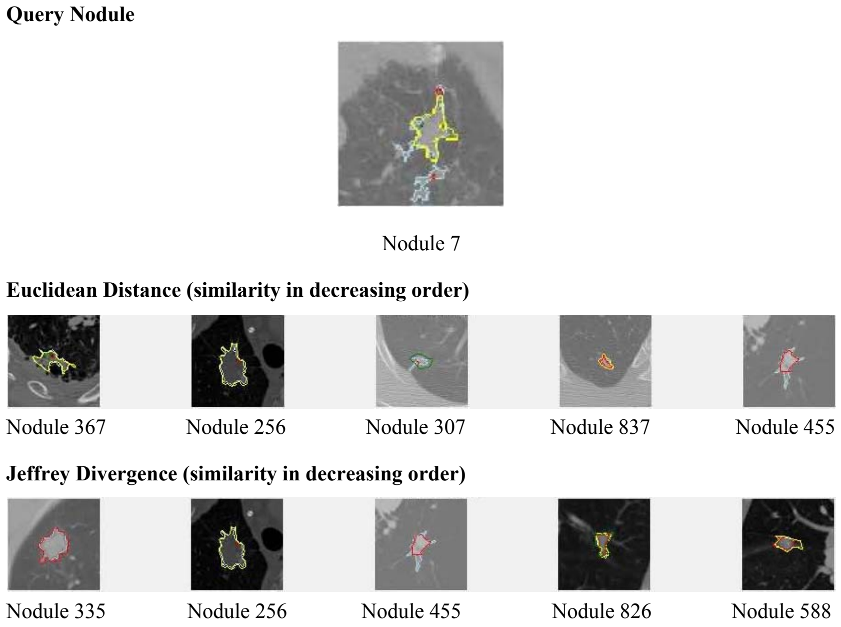

Figure 3: A screenshot of the semantic and content based image retrieval system. Nodule 7 was used as a query image. Nodule 256 and Nodule 455 were retrieved by both content and semantic based methods. This suggests that a relationship between the two retrieval systems exists, but it is still a challenge to determine the relationship.

tigated, such as an ensemble of classifiers, or a structure which incorporates radiologists' feedback into the training. Other future work should include using three dimensional features as opposed to the two dimensional features used in this study.

6. ACKNOWLEDGMENTS

This project was completed with the National Science Foundation award IIS-0755407. We would like to thank Dmitriy Zinovev for the low-level image feature implementations.

7. REFERENCES

[1] S. G. Armato, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, and et al. Lung image database consortium: Developing a resource for the medical imaging research community. *Radiology*, 232:739–748, 2004.

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

[3] N. Asada, K. Doi, and H. Macmahon. Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung disease: pilot study. *Radiology*, 177:857–860, 1990.

[4] K. Ashizawa, T. Ishida, H. Macmahon, C. Vyborny, S. Katsuragawa, and K. Doi. Artificial neural networks in chest radiography: application to the differential diagnosis of interstitial lung disease. *Academic Radiology*, 6:2–9, 1999.

[5] A. S. Barb, C. Shyu, and Y. P. Sethi. Knowledge representation and sharing using visual semantic modeling for diagnostic medical image databases. *IEEE Transaction Information Technology in Biomedicine*, 9, 2005.

[6] G. Dasovich, R. Kim, J. Furst, and D. Raicu. An investigation into the relationship between semantic and content based similarity using lidc. *SPIE Medical Imaging Conference*, February 2010.

[7] A. Fukushima, K. Ashizawa, and T. Yamaguchi. Application of an artificial neural network to high-resolution ct: usefulness in differential diagnosis of diffuse lung disease. *American Journal of Roentgenology*, 183:297–305, 2004.

[8] M. Huber, M. Nagarajan, G. Leinsinger, L. Maximilians, L. Ray, and A. Wismueller. Classification of interstitial lung disease patterns with topological texture features. *SPIE Medical Imaging*, February 2010.

[9] S. A. Jabon, D. S. Raicu, and J. D. Furst. Content-based versus semantic-based similarity retrieval: a LIDC case study. *SPIE Medical Imaging Conference*, February 2009.

[10] C. Kahn, D. Channin, and D. Rubin. An ontology for PACS integration. *Journal of Digital Imaging*, 19(4):316–327, 2006.

[11] M. Lam, T. Disney, M. Pham, D. Raicu, and J. Furst. Content-based image retrieval for pulmonary computed tomography nodule images. *SPIE Medical Imaging Conference*, February 2007.

[12] S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo. Cancer statisitcs 2000. *Ca-Cancer Journal for Clinicians*, 50:7–33, 2000.

[13] Q. Li, F. Li, J. Shiraishi, S. Katsuragwa, S. Sone, and K. Doi. 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.

[14] H. Liu, D. Song, S. M. Rüger, R. Hu, and V. S. Uren. Comparing dissimilarity measures for content-based image retrieval. In *AIRS*, pages 44–50, 2008.

[15] P. Melville and R. Mooney. Constructing diverse classifier ensembles using artificial training examples. *Proceedings of 18th International Joint Conferences on Artificial Intelligence*, pages 505–510, 2003.

[16] L. Monnier-Cholley, H. MacMahon, S. Katsuragawa, J. Morishita, T. Ishida, and K. Doi. Computer-aided diagnosis for detection of interstitial opacities on chest radiographs. *American Journal of Roentgenology*, 171:1651–1656, 1998.

[17] C. Muramatsu, Q. Li, R. Schmidt, J. Shiraishi, and K. Doi. Investigation of psychophysical similarity measures for selection of similar images in the diagnosis of clustered microcalcifications on mammograms. *Medical Physics*, 35:5695–5702, 2008.

[18] C. Muramatsu, Q. Li, R. Schmidt, J. Shiraishi, and K. Doi. Determination of similarity measures for pairs of mass lesions on mammograms by use of bi-rads lesion descriptors and image features. *Academic Radiology*, 16:443–449, 2009.

[19] D. Raicu, D. Zinovev, J. D. Furst, and E. Varutbangkul. Semi-supervised learning approaches for predicting lung nodules semantic characteristics. *Intelligent Decision Technologies Journal*, 3(2), 2009.

[20] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

[21] R. Samala, W. Moreno, Y. You, and W. Qian. A novel approach to nodule feature optimization on thin section thoracic ct. *Academic Radiology*, 16:418–427, 2009.

[22] J. Wang, F. Li, and Q. L. and. Usefulness of texture features for segmentation of lungs with severe diffuse interstitial lung disease. *SPIE Medical Imaging*, February 2010.

APPENDIX

A. ALL 9 SEMANTIC CHARACTERISTICS AND POSSIBLE RATINGS

The table is presented in the next page.

Characteristic	Description	Possible Ratings
Calcification	Calcification appearance in the nodule	1. Popcorn 2. Laminated 3. Solid 4. Non-central 5. Central 6. Absent
Internal Structure	Expected internal composition of the nodule	1. Soft Tissue 2. Fluid 3. Fat 4. Air
Lobulation	Whether lobular shape is apparent from margin or not	1. Marked 2. · 3. · 4. · 5. None
Malignancy	Likelihood of malignancy	1. Highly Unlikely 2. Moderately Unlikely 3. Indeterminate 4. Moderately Suspicious 5. Highly Suspicious
Margin	How well defined the margins are	1. Poorly Defined 2. · 3. · 4. · 5. Sharp
Sphericity	Dimensional shape in terms of roundness	1. Linear 2. · 3. Ovoid 4. · 5. Round
Spiculation	Degree of exhibition of spicules	1. Marked 2. · 3. · 4. · 5. None
Subtlety	Contrast between nodule and surroundings	1. Extremely Subtle 2. Moderately Subtle 3. Fairly Subtle 4. Moderately Obvious 5. Obvious
Texture	Internal density of nodule	1. Non-Solid 2. · 3. Part Solid (Mixed) 4. · 5. Solid