EXPANDING DIAGNOSTICALLY LABELED DATASETS USING CONTENT-BASED IMAGE RETRIEVAL

Anne-Marie Giuca, Kerry A. Seitz Jr.†, Jacob Furst‡, Daniela Raicu‡*

*Pomona College
Claremont, CA, USA
anne-marie.giuca@pomona.edu

†Trinity University
San Antonio, TX, USA
kseitz@trinity.edu

‡DePaul University
Chicago, IL, USA
{jfurst, draicu}@cdm.depaul.edu

ABSTRACT

In computer-aided diagnosis (CAD), having an accurate ground truth is critical. However, the number of databases containing medical images with diagnostic information is limited. Using pulmonary computed tomography (CT) scans, we develop a content-based image retrieval (CBIR) approach to exploit the limited amount of diagnostically labeled data in order to annotate unlabeled images with diagnoses. By applying this CBIR method iteratively, we expand the set of diagnosed data available for CAD systems. We evaluate the method by implementing a CAD system that uses undiagnosed lung nodules as queries and retrieves similar nodules from the diagnostically labeled dataset. In calculating the precision of this system, radiologist- and computer-predicted malignancy data are used as ground truth for the undiagnosed query nodules. Our results indicate that CBIR expansion is an effective method for labeling undiagnosed images in order to improve the performance of CAD systems.

***Index Terms*—**Computer-aided diagnosis, biomedical imaging, semi-supervised learning, cancer detection

1. INTRODUCTION

Lung cancer accounts for the highest number of cancer deaths in the United States each year [1]. Computed tomography (CT) scans can assist radiologists in early detection of lung nodules, which increases the likelihood of a patient's survival [2]. In order to improve lung nodule detection, computer-aided diagnosis (CAD) is effective as a second opinion for radiologists in clinical settings [3]. A dataset with ground truth diagnosis information is essential for CAD systems in order to analyze new cases.

The pulmonary CT scans used in this study were obtained from the Lung Image Database Consortium (LIDC) [4], and we refer to the nodules in this dataset as the LIDC Nodule Dataset. Recently, diagnosis data for some of the nodules were released by the LIDC; however, because the diagnosis was done on a patient, not nodule, level, only the diagnoses belonging to patients with a single nodule could be reliably matched with the nodules in the LIDC Nodule Dataset, resulting in 18 diagnosed nodules (eight malignant, nine benign, and one unknown). The 17 nodules with known diagnoses comprise the initial Diagnosed Subset. Since the diagnoses in the LIDC Diagnosis Dataset are the closest thing to a ground truth available for the malignancy of the LIDC nodules, our goal is to expand the Diagnosed Subset by adding nodules similar to those already in the subset.

To identify these similar nodules and to predict their diagnoses, content-based image retrieval (CBIR) is employed. Increasing the number of nodules for which a diagnostic ground truth is available is important for future CAD applications of the LIDC database.

With the aid of similar images, radiologists' diagnoses of lung nodules in CT scans can be significantly improved [5]. Having diagnostic information for medical images is an important tool for datasets used in clinical CBIR [6]; however, any CAD system would benefit from a larger Diagnosed Subset, since the increased variability in this set would result in more accurately predicted diagnoses for new patients.

1.1. State of the art

Only a limited number of CAD studies have used a pathologically confirmed diagnostic ground truth, since there are few publically available databases with pathological annotations [7]. Nakamura et al. investigated the challenge of distinguishing malignant from benign lung nodules using an artificial neural network (ANN) [8], and they found that their ANN trained on image features outperformed the radiologists. These results, which were validated with a pathological diagnostic ground truth, suggest that radiologists could benefit from the use of their proposed CAD system. Muramatsu et al. [9] used an ANN to learn semantic similarity from content-based features for mammograms from the Digital Database for Screening Mammography [10], an extensive database containing diagnosis data confirmed by pathology.

In CAD applications for which pathological diagnosis data is absent, determining a ground truth is more challenging. In exploring the relationship between content-based similarity and semantic-based similarity for LIDC images, Jabon et al. found that there is a high correlation between image features and radiologists' semantic ratings [11]. Despite this correlation, radiologist malignancy ratings cannot be considered a valid ground truth due to the variability among radiologists [7].

In the absence of diagnostic information, labels can be applied to unlabeled data using semi-supervised learning (SSL) approaches. In SSL, unlabeled data is exploited to improve learning when the dataset contains an insufficient amount of labeled data [12]. Blum and Mitchell pioneered a SSL technique known as co-training for datasets containing large amounts of unlabeled data [13]. Predictors from two independent classifiers were applied to unlabeled data, thereby expanding the training data for the other classifier. CBIR can be used as a machine learning process that trains a system to classify images as relevant or

irrelevant to the query [14]. In a manner similar to co-training, Zhou et al. implemented a semi-supervised active image retrieval (SSAIR) system integrated with relevance feedback to classify unlabeled images from the COREL database with two independent learners [15]. The images labeled as most relevant and irrelevant from each round of relevance feedback were passed to the other learner for re-training.

The limitation of co-training is that one must have two "sufficient and redundant views" for the labeled data, or two sets of information that are sufficient for learning and conditionally independent [13]. In the absence of this information, Szummer and Jakkola used a Markov random walk approach to label data in low dimensional datasets, where a kNN graph was constructed and edge weights were assigned based on Euclidean distance [15].

In the current study, we adopted a semi-supervised approach for labeling undiagnosed nodules in the LIDC. CBIR was used to label nodules most similar to the query with respect to Euclidean distance of image features. By evaluating the method with a CAD application, we determined how to effectively expand the Diagnosed Subset with CBIR.

2. METHODS

2.1. The Datasets

The LIDC database, released in 2009, contains 399 pulmonary CT scans. Up to four radiologists analyzed each scan by identifying nodules and rating the malignancy of each nodule on a scale of one to five [4]. To reduce the variability among radiologists, the mode of the radiologists' ratings was used [11]. Nodules with malignancy ratings of one or two were considered benign, four or five were malignant, and three were unknown. Each nodule was represented by one slice [16], and 63 image features were extracted for each nodule based on texture, size, shape, and intensity [17]. The three feature extraction methods used to obtain these 63 features from the LIDC images were Haralick co-occurrence, Gabor filters, and Markov random fields [17]. The number of nodules was reduced to 914 by removing nodules smaller than five-by-five pixels because features extracted from these smaller nodules are imprecise.

For each nodule, computer-predicted probability distributions for malignancy were obtained using a CAD algorithm described in previous work [18]. For each malignancy rating (one to five), a probability was assigned based on the predictions of an ensemble of classifiers trained using radiologists' ratings and constructed from the 63 image features. For this study, each nodule was then assigned a computer-predicted malignancy rating of malignant, benign, or unknown based on its probability distribution.

The recently released LIDC Diagnosis Dataset assigned a malignant, benign, or unknown diagnosis (based on biopsy, surgical resection, progression or response, a review of radiological images showing two years of stability, or an unknown method) to some of the patients in the LIDC Nodule Dataset [4]. Of the 914 nodules in the LIDC Nodule Dataset, only 17 nodules (eight malignant and nine benign) could confidently be assigned a diagnosis from the LIDC Diagnosis Dataset. This set of 17 nodules and the nodules subsequently added to this set will be referred to as the Diagnosed Subset.

2.2. Candidate Identification

CBIR was used to identify candidate nodules from the LIDC Nodule Dataset for which a diagnosis could be assigned (Figure 1).

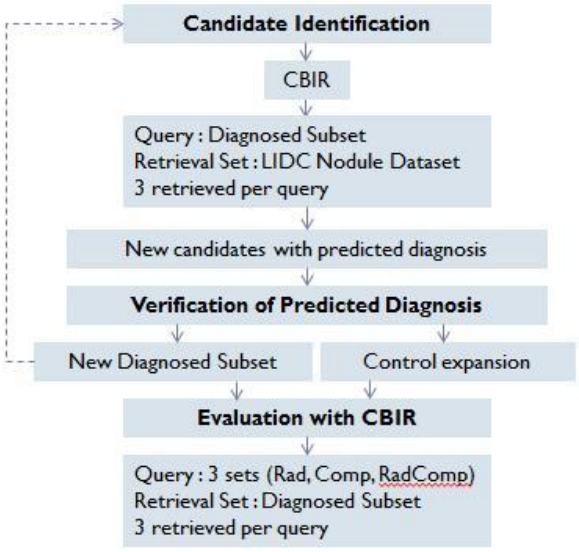

Figure 1. Summary of CBIR method of expanding the Diagnosed Subset; CBIR expansion occurs iteratively.

For each nodule in the LIDC Nodule Dataset, the 63 image features were used to calculate Euclidean distance between nodule pairs, representing their similarity. Each nodule in the Diagnosed Subset was then used as a query to retrieve the three most similar images from the remaining nodules in the LIDC Nodule Dataset. The retrieved nodules were assigned predicted malignancy ratings based on the query nodule used to retrieve it (e.g., the three nodules retrieved by a malignant query were assigned malignant diagnoses). If both a malignant query and a benign query retrieved the same nodule, that nodule was disregarded. The newly identified nodules were considered candidates for addition to the Diagnosed Subset.

2.3. Verification of Predicted Diagnosis

Nodules to be added to the Diagnosed Subset were selected from the candidates described above. If a candidate nodule belonged to a patient in the LIDC Diagnosis Dataset and the nodule's predicted diagnosis based on CBIR agreed with the pathologically-determined patient-level diagnosis, then that nodule was added to the Diagnosed Subset. By verifying the predicted diagnosis with the pathologically-determined diagnosis, this process guarantees the accuracy of the CBIR-based diagnostic labeling.

2.4. Diagnosed Subset Evaluation

In order to evaluate the CBIR expansion method, an independent CBIR system was implemented mimicking a potential CAD application. In this CAD system, a radiologist would use an undiagnosed nodule as a query against a retrieval set (the set of nodules from which the query set retrieves images) of diagnosed nodules to aid in diagnosing a newly scanned nodule [7].

2.4.1. Query and Retrieval Sets

In this CAD scenario, the Diagnosed Subset was used as the retrieval set, and the malignancy for this set was determined by the method described in sections 2.2 and 2.3. This set was balanced to contain an equal number of malignant and benign nodules so that neither rating was more likely to be retrieved randomly.

Three different "undiagnosed" query sets containing subsets of the LIDC Nodule Dataset were used, since neither computer-predicted nor radiologist-predicted malignancy ratings can be considered ground truth due to high variability between radiologists' ratings [7]. Each of these query sets differed in diagnostic ground truth. The first query set (Rad) used the radiologist-predicted malignancy, the second set (Comp) used the computer-predicted malignancy, and the third set (RadComp) used only those nodules for which the radiologist- and computer-predicted malignancies agreed. For each query set, nodules with unknown malignancies were removed, and the set was balanced to contain an equal number of malignant and benign nodules. The radiologist-predicted, computer-predicted, and radiologist-computer-agreement query sets contained 268, 216, and 148 nodules, respectively, after these modifications.

2.4.2. Precision

Using the query and retrieval sets as described above, average precision after 3, 5, 10, and 20 images retrieved was calculated. A retrieved nodule was considered relevant if its diagnosis matched the malignancy rating (either radiologist-predicted, computer-predicted, or both) of the query nodule. Initial precision values were obtained by using the 17 nodules in the initial Diagnosed Subset as the retrieval set. Then, nodules were added to this set as described in sections 2.2 and 2.3. Precision was recalculated, and the nodule addition process was repeated iteratively using the new Diagnosed Subset. In each subsequent iteration, only the newly added nodules in the Diagnosed Subset were used to identify new candidates. This process repeated until no candidate nodules were added to the Diagnosed Subset following an iteration.

It is important to note that the CBIR directed expansion method used the Diagnosed Subset as the query set and the LIDC Nodule Dataset as the retrieval set. In contrast, the CBIR CAD application used subsets of the LIDC Nodule Dataset as the query set and the Diagnosed Subset as the retrieval set for the precision calculations. Thus, although CBIR was involved in both the identification of candidate nodules and the evaluation of the selected nodules, these two processes were not identical.

2.4.3. Control Group

In order to test whether or not changes in precision with the expanded Diagnosed Subset were related only to the increasing size of the dataset, a control experiment was also performed. The control set contained the 17 nodules from the initial Diagnosed Subset, and additional nodules were added by randomly selecting nodules from the LIDC Nodule Dataset if they were not already in the control set. The numbers of malignant and benign nodules added at each iteration by this random expansion method were equal to the numbers of malignant and benign nodules added by the CBIR expansion method at that iteration. Precision was calculated as described in sections 2.4.1 and 2.4.2, using the control set as the retrieval set. The radiologist predicted malignancy was used as ground truth for the control group in precision calculations because in preliminary testing, this malignancy assignment produced results that differed least from the CBIR expansion method and was thus the most rigorous test.

2.5 Multiple Linear Regression

In addition to CBIR, another independent method of expanding the Diagnosed Subset was tested. In this approach, a multiple linear regression (MLR) model for malignancy was constructed based on the 63 extracted image features for the nodules in the Diagnosed

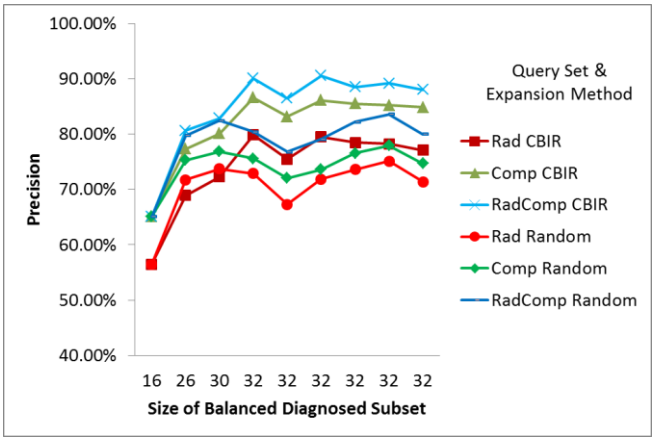

Figure 2. Average precision after three images retrieved using the CBIR and control methods to expand the Diagnosed Subset. Each point on the x-axis represents a discrete iteration of expansion.

Subset. This model was then used to predict diagnoses for the remaining nodules in the LIDC Nodule Dataset. The 20 most likely malignant and 20 most likely benign nodules were assigned malignant and benign diagnoses, respectively. These nodules comprised the set of newly identified candidates to add to the Diagnosed Subset. Forty candidate nodules were identified because preliminary testing showed that approximately 40 candidates were identified using the CBIR method, allowing us to more directly compare the two methods. As with CBIR expansion, the MLR expansion method was compared to a control method that randomly selected nodules for addition to the Diagnosed Subset.

3. RESULTS

The average precision results obtained while expanding the Diagnosed Subset using CBIR are presented in Figure 2. Only the results for precision after 3 images retrieved are shown because the size of the balanced Diagnosed Subset remained relatively small throughout the expansion process, and retrieving a large number of images from a small retrieval set can result in misleadingly low precision values. Although the size of the balanced Diagnosed Subset remained constant between some iterations, nodules were still being added to this set. However, the malignancy of these added nodules was the same as the predominant malignancy rating in the Diagnosed Subset, so the limiting factor for balancing remained constant. One-tailed t-tests with $p<0.05$ were used to compare the control expansion to CBIR expansion for each of the query sets. CBIR expansion resulted in significantly higher precisions than control expansion after the balanced Diagnosed Subset reached a size of 32. These results indicate that CBIR is a reliable method of expanding labels to undiagnosed images. Validation of these results was ensured by only adding diagnostic labels to images for which the diagnosis could be confirmed by information in the LIDC Diagnosis Dataset. Further validation was provided by measuring precision of the proposed CAD system with three different diagnostic ground truths (radiologist-, computer-, and radiologist and computer-predicted malignancy).

Evaluating precision obtained with the MLR expansion method revealed that the method was outperformed by both the control and CBIR expansion methods. These results indicate that MLR is an ineffective approach to expanding diagnostic labels. MLR selects nodules for addition to the Diagnosed Subset based on a malignancy model developed from all the nodules, whereas CBIR expansion adds nodules that are similar to specific nodules in the Diagnosed Subset. Therefore, it is possible that CBIR's more

direct comparison of nodules is more effective in applying diagnostic labels.

4. CONCLUSION

CBIR is an effective method for expanding the Diagnosed Subset by labeling nodules which do not have associated diagnoses. This method outperforms control expansion, yielding higher precision values when tested with a potential CAD application that requires a diagnostically accurate ground truth. By increasing the size of the Diagnosed Subset from 17 to 74 nodules, CBIR expansion provides greater variability in the retrieval set, resulting in retrieved nodules that are more similar to undiagnosed queries. The proposed CBIR expansion method can be applied to other image databases containing large quantities of unlabeled data with few labeled instances. An expanded set of diagnosed images is also useful for non-CBIR CAD systems, which require large datasets for robust and unbiased training and testing. In future studies, we will investigate using different distance metrics for nodule similarity when identifying candidates with the CBIR expansion method. We also plan to implement an expansion method using decision trees to identify candidate nodules to add to the Diagnosed Subset.

5. ACKNOWLEDGMENTS

This work was supported in part by NSF award 1062909.

6. REFERENCES

[1] R. Siegel, E. Ward, O. Brawley, and A. Jemal, "Cancer statistics, 2011: The impact of eliminating socioeconomic and racial disparities on premature cancer deaths," *CA: A Cancer Journal for Clinicians*, Wiley-Blackwell, vol. 61, pp. 212-236, 2011.

[2] C. I. Henschke, D. I. McCauley, D. F. Yankelevitz, D. P. Naidich, G. McGuinness, O. S. Miettinen, D. M. Libby, M. W. Pasmantier, J. Koizumi, N. K. Altorki, and J. P. Smith, "Early Lung Cancer Action Project: overall design and findings from baseline screening," In Porter, J. C. and Spiro, S. G., "Detection of early lung cancer," *Thorax*, vol. 55 (sump 1), pp. S56-S62, 2000.

[3] D. Wormanns, M. Fiebich, M. Saidi, S. Diederich, and W. Heindel, "Automatic detection of pulmonary nodules at spiral CT: clinical application of a computer-aided diagnosis system," *European Radiology*, vol. 12, pp. 1052-1057, 2002.

[4] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. van Beek, D. Yankelevitz, et al., "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, pp. 915-931, 2011.

[5] 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*, vol. 30, pp. 2584–2593, 2003.

[6] H. Müller and J. Kalpathy-Cramer, "Putting the Content Into Context: Features and Gaps in Image Retrieval," In J. Tan, *New Technologies for Advancing Healthcare and Clinical Practices*, IGI Global, Hershey PA, pp. 105-115, 2011.

[7] W. H. Horsthemke, D. S. Raicu, J. D. Furst, and S. G. Armato III, "Evaluation Challenges for Computer-Aided Diagnostic Characterization: Shape Disagreements in the Lung Image Database Consortium Pulmonary Nodule Dataset," In J. Tan, *New Technologies for Advancing Healthcare and Clinical Practices*, IGI Global, Hershey PA, pp. 18-43, 2011.

[8] K. Nakamura, H. Yoshida, and R. Engelmann, "Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks," *Radiology*, vol. 214, pp. 823-830, 2000.

[9] 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*, vol. 35, pp. 5695-5702, 2008.

[10] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. Kegelmeyer, "The digital database for screening mammography," *Fifth International Workshop on Digital Mammography*, pp. 212-218, 2001.

[11] 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*, Orlando, February 2009.

[12] Z.-H. Zhou, "Learning with Unlabeled Data and Its Application to Image Retrieval," *PRICAI'06 Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence*, 2006.

[13] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT '98)*, pp. 92-100, 1998.

[14] Z.-H. Zhou, K.-J. Chek, and Y. Jiang, "Exploiting unlabeled data in content-based image retrieval," *Proceedings of the 15th European Conference on Machine Learning*, pp. 525–536, 2004.

[15] M. Szummer and T. Jaakkola, "Partially labeled classification with Markov random walks," *Advances in Neural Information Processing Systems*, pp. 945-952, 2002.

[16] R. Kim, G. Dasovich, R. Bhaumik, R. Brock, J. D. Furst, and D. S. Raicu, "An Investigation into the Relationship between Semantic and Content Based Similarity using LIDC", *ACM International Conference on Multimedia Information Retrieval (MIR) 2010*, Philadelphia, March 2010.

[17] 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*, San Diego, February 2007.

[18] D. Zinovev, D. Raicu, J. Furst, and S. Armato III, "Predicting Radiological Panel Opinions using a Panel of Machine Learning Classifiers," *Algorithms*, vol. 2, pp. 1473-1502, 2009.