Expanding Diagnostically Labeled Datasets Using Content-Based Image Retrieval

[1]Anne-Marie Giuca, [2]Kerry A. Seitz, Jr., [3]Jacob D. Furst, [3]Daniela S. Raicu

[1]Pomona College, 333 North College Way, Claremont, California, USA 91711

[2]Trinity University, One Trinity Place, San Antonio, Texas, USA 78212

[3]DePaul University, 243 South Wabash Avenue, Chicago, Illinois, USA 60604

Introduction

Pulmonary computed tomography (CT) scans assist radiologists in early detection of lung nodules, and computer-aided diagnosis (CAD) is an effective second opinion for radiologists. For CAD systems, having a diagnostic ground truth is necessary; however, only a limited number of medical image databases contain diagnostic labels. We demonstrate that content-based image retrieval (CBIR) is an effective tool for annotating unlabeled images with diagnoses.

Datasets

- Lung Image Database Consortium (LIDC): 399 CT scans with nodule malignancy ratings from up to 4 radiologists
- LIDC Nodule Dataset: 914 nodules, each annotated with radiologist-predicted and computer-predicted malignancy ratings
- LIDC Diagnosis Dataset: 17 nodules matched to patient-level diagnosis (9 benign, 8 malignant)
- Diagnosed Subset: includes LIDC Diagnosis Dataset and nodules that are subsequently labeled

Figure 1. CT image from the LIDC.

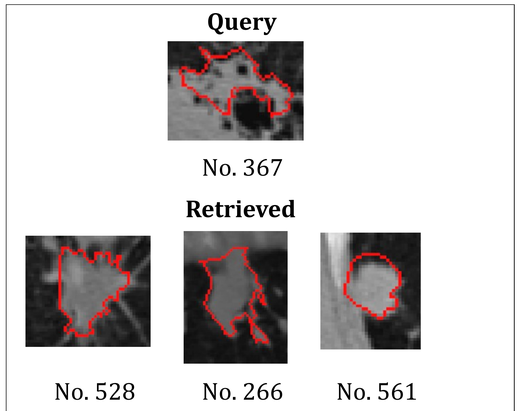

Figure 2. Example of CBIR with the LIDC Nodule Dataset. The query (top) was used to retrieve the three most similar nodules (bottom). The nodule number denotes identification within the LIDC.

Methods

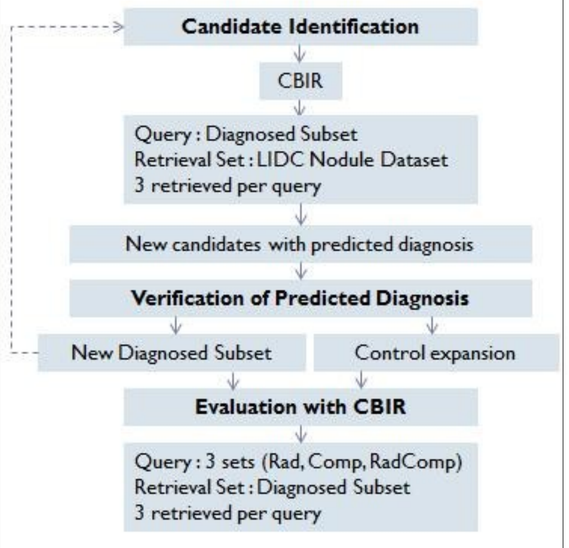

Figure 3. Summary of CBIR method of expanding the Diagnosed Subset; CBIR expansion occurs iteratively.

Results

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.

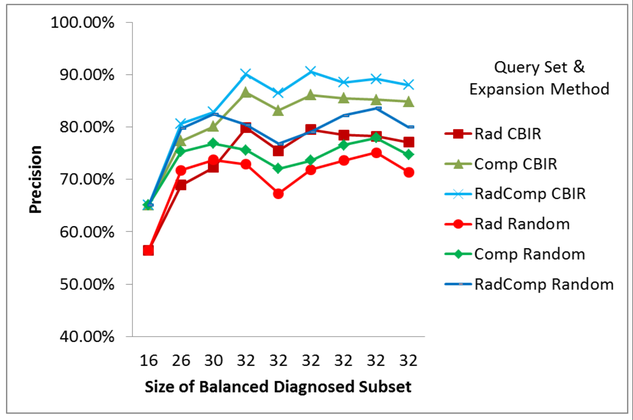

Figure 4. Average precision after three images retrieved using the CBIR and control methods to expand the Diagnosed Subset. Each point on the x-axis represents a discrete iteration of expansion.

Acknowledgments

This work was supported in part by NSF award 1062909.

Conclusion

CBIR is an effective method for expanding diagnostically labeled datasets. 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.

References

- S. G. Armato III, 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.
- 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.
- 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.
- 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.