Multi-level Ground Glass Nodule Detection and Segmentation in CT Lung Images

Yimo Tao[1,2], Le Lu[1], Maneesh Dewan[1], Albert Y. Chen[1,3], Jason Corso[3], Jianhua Xuan[2], Marcos Salganicoff[1], and Arun Krishnan[1]

[1] CAD R&D, Siemens Healthcare, Malvern, PA USA
[2] Dept. of Electrical and Computer Engineering, Virginia Tech, Arlington, VA USA
[3] Dept. of CSE, University at Buffalo SUNY, Buffalo, NY USA

Abstract. Early detection of Ground Glass Nodule (GGN) in lung Computed Tomography (CT) images is important for lung cancer prognosis. Due to its indistinct boundaries, manual detection and segmentation of GGN is labor-intensive and problematic. In this paper, we propose a novel multi-level learning-based framework for automatic detection and segmentation of GGN in lung CT images. Our main contributions are: firstly, a multi-level statistical learning-based approach that seamlessly integrates segmentation and detection to improve the overall accuracy for GGN detection (in a subvolume). The classification is done at two levels, both voxel-level and object-level. The algorithm starts with a three-phase voxel-level classification step, using volumetric features computed per voxel to generate a GGN class-conditional probability map. GGN candidates are then extracted from this probability map by integrating prior knowledge of shape and location, and the GGN object-level classifier is used to determine the occurrence of the GGN. Secondly, an extensive set of volumetric features are used to capture the GGN appearance. Finally, to our best knowledge, the GGN dataset used for experiments is an order of magnitude larger than previous work. The effectiveness of our method is demonstrated on a dataset of 1100 subvolumes (100 containing GGNs) extracted from about 200 subjects.

1 Introduction

Ground Glass Nodule(GGN) is a hazy area of increased attenuation in CT lung images, often indicative of bronchioloalveolar carcinoma (BAC) [1], that does not obscure underlying bronchial structures or pulmonary vessels. These faint pulmonary nodules are reported to have a higher probability of becoming malignant than solid nodules [1]. Hence early detection [2] and treatment of GGN are important for improving the prognosis of lung cancer. Furthermore, recent studies have shown that tracking the growth pattern of GGNs is informative and useful for quantifying and studying the progress of diseases over time [3]. Therefore, it is highly desirable to have algorithms that not only detect the GGN but are also capable of segmenting the GGN with good accuracy. However, due to their indistinct boundaries and similarity to its surrounding structures, consistent labeling

G.-Z. Yang et al. (Eds.): MICCAI 2009, Part II, LNCS 5762, pp. 715–723, 2009.

of GGN at voxel-level is difficult for both computers and radiologists, with high inter- and intra-person errors [4]. On the other hand, the appearances of GGN on CT images, such as its shape, pattern and boundary, are very different from solid nodules. Thus, algorithms developed exclusively for solid nodule segmentation are likely to produce inaccurate results when directly applied to GGN. [5] addresses a GGN segmentation method using Markov random field representation and shape analysis based vessel removal method, but no GGN detection was exploited in this approach. [4] adopts the probability density functions (PDF) for modeling and segmenting GGN and PDFs are shown to be valid of distinguishing GGN and other lung parenchyma. An interactive 2D semi-automatic segmentation scheme is proposed in [6], which allows measuring the pixel opacity value of the GGN quantitatively, by constructing a graph Laplacian matrix and solving a linear equation system. This may be quite labor-intensive for radiologists to go through slices in order to obtain the final GGN nodule boundaries and opacity values.

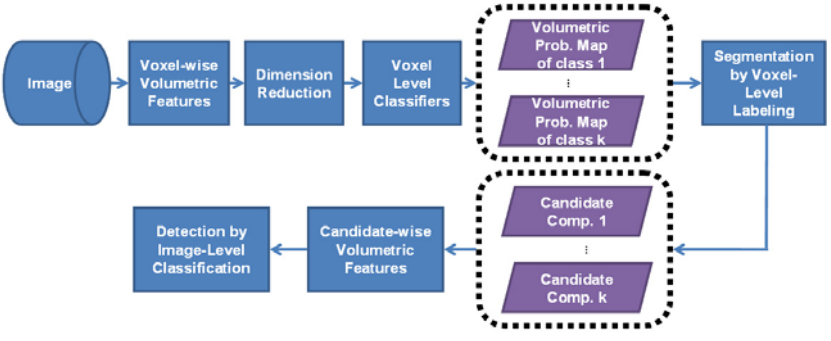


Fig. 1. Flow chart of the proposed approach

In this work, we propose a novel multi-level learning based approach for automatic segmentation and detection for GGN. Our method is composed of two parts as shown in Fig. 1: (1) voxel-wise soft labeling/segmentation, and (2) object-scale image classification/detection. The voxel-level GGN and non-GGN class labeling and segmentation works as follows: given a candidate sub-volume, our system will label each voxel with the probability likelihood of whether it comes from a "healthy" (non-GGN) or "diseased" (GGN) tissue by examining a 3D sliding window centered at that voxel; this labeled or probability weighted image can be viewed as a segmentation map of the GGN region in the original image. After this, the object-level image classification/detection module takes the label map along with other raw information and then classifies the whole sub-volume as positive (containing GGN) or negative. Thus only if the whole sub-volume is classified as positive or "diseased", we will conclude that detection has occurred. Although the method is designed for combined segmentation and detection of GGNs, it can be generalized to other medical imaging problems as well. The presented multi-level probabilistic aggregation process is partially motivated by earlier work in natural scene recognition and object detection [7,8].

The other key aspects of our work is that we use an comprehensive set of features to capture GGN appearance, and the experimental evaluation is done on a much larger dataset compared to previous studies. Note that our algorithm works at a sub-volume level, and it assumes that a candidate generation or interest point detection algorithm is run on a CT volume to obtain locations (typically hundreds) around which isotropic sub-volumes of fixed size (60×60×60 voxels) are extracted. This is typical in a computer-aided detection (CAD) system. In this paper, we focus on detection and segmentation at this sub-volume (SV) level, which is an integral and important part of the CAD system.

2 Methods

2.1 Voxel-Wise Volumetric Features

Our system computes a comprehensive collection of gray-level and texture features from a cubic sub-volume of interest (sVOIs) of size 7×7×7 voxels across the larger sub-volume (SV). These features are briefly described below:

Gray Level Co-occurrence Matrix (GLCM) [9] is widely used for analyzing texture of 2D image. The co-occurrence matrix stores the co-occurrence frequencies of the pairs of gray levels, which are configured by a distance d and orientation o. Its extension to 3D cases is also practicable, as shown in [10]. The 3D method directly searches for gray level pairs in 26 directions on multiple planes to construct the co-occurrence matrix, whereas the 2D method exploits 8 directions in a single 2D plane. We then extract eight features from the constructed GLCM, including energy, entropy, correlation, inverse difference moment, inertia, cluster shade, cluster prominence, and Haralick correlation [9].

Local Binary Pattern (LBP) is an intensity- and rotation-invariant generalization of the GLCM method. We employ the volumetric LBP-Top [11] technique, an extension of the two-dimensional LBP operator, for parenchymal texture analysis in CT images.

Wavelets are another important and commonly used feature descriptor for texture analysis, due to their effectiveness in capturing localized spatial and frequency information and multi-resolution characteristics. Here, we extract mean intensities in the decomposed eight bands using 3D Harr wavelet. **Vesselness** and **Blobness**, computed based on the eigen-analysis of hessian matrix, have also been employed for vascular or blob-like structure detection or enhancement. We implement a 3D multi-scale version of Blobness and Vesselness feature extraction module for handling both bright and dark objects. Note that the Wavelets, Vesselness and Blobness depend on their own scales of spatial supporting settings, and the actual neighborhood may be larger or smaller than the size of 7×7×7.

We also extract two groups of first order **gray-level features**, composed of **(i) gray level statistics features**, including minimum, maximum, mean, standard deviation, skewness and kurtosis, and **(ii) pixel index ratios**, including the ratios of low density pixels within (-1024, -950] Hounsfield unit(HU), medium

density values within (-950, -765] HU, and medium-high density values within (-765, -450] HU.

Since the intensity values in the CT scans usually have a large range from -1024 to 1024 HU, texture feature calculation directly on HU values is computationally intensive and sensitive to noise. Therefore, we preprocess images using the multi-level thresholding *Otsu* method [12] to adaptively merge together image regions with similar gray levels. The resulting image is represented by individual texture primitives coded by a smaller gray-level domain. All texture-based features are extracted from this preprocessed image.

2.2 Segmentation by Voxel-Level Labeling

We treat the segmentation of GGN as a probabilistic voxel-level labeling problem. For each input sub-volume (SV), a total 39 volumetric intensity-texture features are calculated for each scanned sVOI of size 7×7×7 voxels. Based on our 3D annotation maps of GGN and Non-GGN, feature vectors are split into positives and negatives and fed into an off-line learning process to train a probabilistic Gaussian mixture density model, in the lower-dimensional feature space after supervised dimension reduction. Finally the classifier takes each SV and produces its corresponding volumetric GGN-class probability map.

For the voxel-level labeling/classification problem, the size of training samples (as scanned volume of size 7×7×7 voxels) can be really large (greater than 100,000). This requires choosing classifiers with good scalability. We choose linear discriminant analysis (LDA) along with Gaussian Mixture Models (GMM) as our classifier, i.e., GMM is used to learn the distribution of the classes in the LDA projected subspace. For each of the binary GGN and Non-GGN class, LDA is first exploited to further project the extracted features into a lower dimension, and GMM consisting of k-Gaussian distributions of different means and variances are then fit according to the training data, using Expectation-Minimization with multiple random initialization trials. Note that, the positive GGN class probability maps (PDM) are sufficient to extract and detect GGNs. The negative probability map is redundant and discarded. Note that, we perform model selection to choose the number k of Gaussian functions using the Bayesian Information criterion (BIC) and value of k is 3 and 5 for positive and negative class respectively. Other functions, e.g., t-distribution can also be explored, but we plan to investigate that in future work.

As there are many different types of tissues inside the CT lung image, such as vessel, airways, and normal parenchymal, the single-layer LDA classifier may have many false positives originating from this multi-tissue background. To reduce these GGN false positives, a multi-phase classification approach is adopted. It starts with the positive class output probability map from single phase, and treat it as a new image. This output image contains for each voxel a probability that it belongs to the structure to be enhanced (GGN). Next, another round of voxel-level feature extraction/selection and LDA-GMM training process is conducted using both the original image and the output image from the previous phase(s). All these intensity-texture features, in the joint image and PDM

domain, are used to train a new classifier. This process can be iterated many times, as a simplified "Auto-Context" [13]. The rationale behind this approach is that the structure to be enhanced will be more distinctive in the (intermediate) enhanced image than in the original image. Therefore adding features from these weighed images will result in potentially more discriminative features between the positive regions and the spurious responses from the previous phase(s).

The multi-phase "Auto-Context" like process not only improves the overall performance but can also be used to speed up supervised enhancement by rejecting well classified training samples from the next phase. A simple classifier (e.g., using very few features) can be used in the first phase(s) to quickly throw away "easy" voxels and only the more difficult voxels are considered in the next phase(s). The classification thresholds on normalized probability values are automatically estimated by setting an operating point on receiver operating characteristic curve (ROC) for high recall and moderate false positive deduction.

2.3 Object-Level Labeling and Detection

At this stage, the goal is to locate and extract GGN candidates on the computed 3D probability map from the previous voxel-labeling step. The multiscale blobness filtering, with a larger smoothing kernel (than the voxel-level step) is used to capture the GGN shape. It is applied on each voxel to obtain a volumetric blobness likelihood map. Then, we multiply the voxel-level probability map with this blobness shape likelihood map to obtain another probability map which we refer to as shape-prior refined probability map (SPM). The SPM helps suppress spurious responses (false positives). The Otsu thresholding method[12] is applied again for discrete quantitization of SPM. We use connected component labeling to obtain disjointed objects as GGN candidates. Simple volume size based rules is used to reduce the number of candidate so that multiple candidates per volume are kept as the inputs to our object level classifier. We also incorporate the position prior information into candidate selection procedure.

For training, the manually annotated GGN segmentation is used to assign labels to the connected component candidates as true or false GGNs. The above-mentioned candidate generation procedure is also applied on negative volumes (without GGN) to obtain more negative samples for training. Given that each GGN candidate is associated with its discrete binary 3D mask, 39 intensity-texture features (mentioned in Section 2.1), are recomputed within this binary supporting region. Note that, these features are not computed on the $7\times7\times7$ window as done earlier for the voxel-level classification stage. Many features are aggregations of local statistics over a spatial neighborhood so that they are size-scalable. In addition, we also calculate the volume size, the sphericity, and the mean on the PDM per candidate to form the final feature vector. For simplicity, we use the same LDA+GMM classifier as in section 2.2 to train GGN/non-GGN connected component candidate detector.

3 Results

Data. We collected total 1100 lung CT subvolumes, including 100 positive samples with GGN and 1000 negative samples without GGN. These subvolumes were randomly sampled from the outputs of a GGN candidate generator algorithm (discussed earlier in Section 1), on 153 healthy and 51 diseased patients. All subvolumes were sampled to produce approximately 0.6mm isotropic voxel resolution. GGN masks on randomly selected 60 positive samples were annotated. The remaining 40 positive subvolumes (with no ground truth of GGN masks) along with 300 randomly selected negative subvolumes were used only for the hold-out performance testing of GGN detection at the final classification stage.

Voxel-scale Classification & Labeling. We further splited these 60 positive subvolumes (with annotated GGN masks) and other 700 negative volumes into two parts for training and testing the voxel scale classification. Voxel scale data samples (as 7×7×7 boxes) were extracted on a 7×7×7 sampling grid, augmented with manually labeled GGN annotation masks. The training dataset had 40 positive and 500 negative subvolumes for the three-phase classifier training in section 2.2 (with likelihood ratio testing threshold settings as 0.05, 0.1 and 0.2 respectively for high recall and moderate false positive reduction). And the remaining subvolumes were used for testing. Table. 1 showed the voxel-level accuracies for the first phase classifier, first two phase classifiers, and all three phase classifiers. It was clearly evident that the further reduction of false positive samples, with increasing classification phases, substantially improved the overall classification performance. The ROC curve for the first phase classifier was shown in Fig. 2, with Area Under Curve (AUC) as 0.9412. We empirically found that the performance is stable with the sVOI size in a range of 5 to 15 voxels (note that the default value is 7).

To measure the level of agreement between human annotations and the segmented GGN (with Otsu thresholding and connected component), the *Jaccard* similarity coefficient (JC) and the volume similarity (VS) were exploited. The JC was defined as:

$$JC = \frac{X \cap Y}{X \cup Y} \tag{1}$$

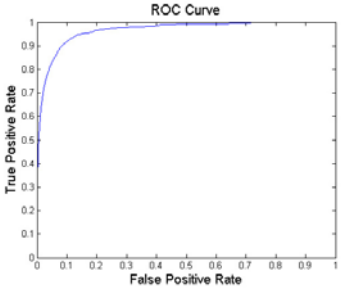


Fig. 2. The top level voxel-level GGN classification ROC curve

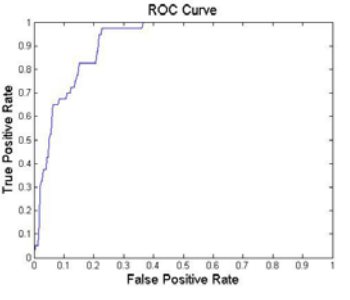
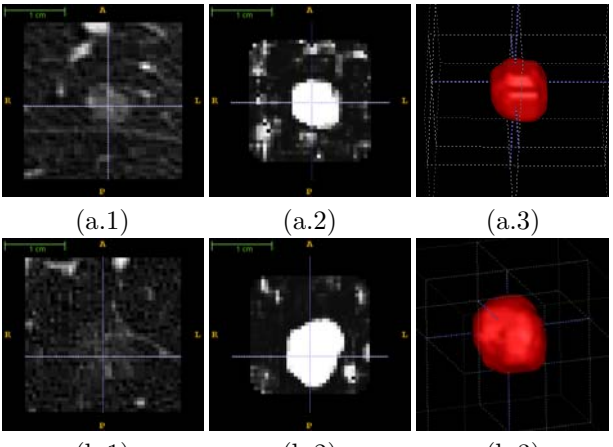


Fig. 3. The object-level classification ROC curve for GGN detection

Table 1. The multi-level voxel scale classification accuracy performance

	First Phase	First Two Phases	All Three Phases
GGN Samples	99.82%	96.62%	89.87%
Negative Samples	56.53%	80.86%	92.93%
Overall	65.37%	84.22%	92.28%



(a.1) (a.2) (a.3)

(b.1) (b.2) (b.3)

Fig. 4. Results the voxel-level classification: (a.1) and (b.1) The original CT images, (a.2) and (b.2) The volumetric probability map produced by voxel-scale classifiers, (a.3) and (b.3) the rendered segmentation of GGN

where JC measured the degree of overlap between two sets: X and Y and JC=1 when the two sets were totally overlapped. The VS was defined as:

$$VS = 1 - \frac{|\,||X|| - ||Y||\,|}{||X|| + ||Y||} \tag{2}$$

where VS measured the degree of similarity in the volume size of two sets X and Y, and the operator $||.||$ denoted the volume size of a set. It was equal to one, when the two sets had equal volumes. We reserved 20 positive GGN subvolumes for testing these two metrics, and used the remaining 40 positive subvolumes with all the negative volumes to train the voxel-level classifier. The average JC coefficient was 0.68, and the average VS was 0.865.

Object-scale Classification & GGN Detection. Fig. 3 showed the object level GGN classification performance with AUC as 0.914. For the hold-out testing set, 33 out of 40 GGN subvolumes were correctly detected, with a false negative rate is about 20%. We believed this result is promising and we planed to explore more descriptive features and other types of classifiers (e.g., SVM, boosting) to further investigate this problem. As compared with [4], the most relevant previous work in which only 10 GGN nodules were used for both training and

testing, our studied GGN dataset is 10 times larger (i.e., 60 for training and 40 for testing). [6] uses 40 2-D CT image slices from 11 patients. Finally, illustrative examples of GGN labeling and segmentation were shown in Fig. 4.

4 Conclusion

In this paper, we presented a novel multi-level learning-based approach for automatic GGN detection that fuses segmentation and detection to improve the overall accuracy. We exploited a comprehensive set of features to encapsulate the GGN appearance. Our approach proposed a two-level classification by first generating a GGN class PDM at a voxel scale, then extracting object scale descriptors from this PDM, and then finally classifying the existence of GGN within CT subvolume candidate. The GGN segmentation (soft) mask was a desired byproduct of our approach. Our method was validated using extensive evaluations on a much larger GGN dataset than previously reported, of 1100 lung CT subvolumes from about 200 patients.

References

1. Henschke, C.I., Yankelevitz, D.F., Mirtcheva, R., McGuinness, G., McCauley, D., Miettinen, O.S.: CT Screening for Lung Cancer Frequency and Significance of Part-Solid and Nonsolid Nodules. Amer. Jour. Roentgenology 178(5) (2002)
2. Godoy, M., Ko, J., Kim, T., Naidich, D., Bogoni, L., Florin, C., Groot, P., White, C., Vlahos, I., Park, S., Salganicoff, M.: Effect of computer-aided diagnosis on radiologists detection performance of subsolid pulmonary nodules on ct: Initial results. In: American Roentgen Ray Society, ARRS (2009)
3. Park, C., Goo, J., Lee, H., Lee, C., Chun, E., Im, J.: Nodular ground-glass opacity at thin-section ct: Histologic correlation and evaluation of change at follow-up. RadioGraphics 27(2), 391–408 (2007)
4. Zhou, J., Chang, S., Metaxas, D., Zhao, B., Ginsberg, M., Schwartz, L.: Automatic detection and segmentation of ground glass opacity nodules. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 784–791. Springer, Heidelberg (2006)
5. Zhang, L., Fang, M., Naidich, D., Novak, C.: Consistent interactive segmentation of pulmonary ground glass nodules identified in ct studies. In: SPIE Medical Imaging (2004)
6. Zheng, Y., Kambhamettu, C., Bauer, T., Steiner, K.: Estimation of ground-glass opacity measurement in CT lung images. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5242, pp. 238–245. Springer, Heidelberg (2008)
7. Lu, L., Toyama, K., Hager, G.: A two level approach for scene recognition. In: IEEE Conf. CVPR, pp. 688–695 (2005)
8. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. Int. J. of Computer Vision 77, 259–289 (2008)
9. Haralick, R.: Statistical and structural approaches to texture. Proceedings of the IEEE 67, 786–804 (1979)

10. Xu, Y., Sonka, M., MnLenan, G., Guo, J., Hoffman, E.: MDCT-based 3-d texture classification of emphysema and early smoking related lung pathologies. IEEE Transactions on Medical Imaging 25(4) (2006)
11. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6) (2007)
12. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics 9, 62–66 (1979)
13. Tu, Z.: Auto-context and its application to high-level vision tasks. In: IEEE Conf. CVPR, pp. 1–8 (2008)