Computer-Aided Detection of Ground Glass Nodules in Thoracic CT Images Using Shape, Intensity and Context Features

Colin Jacobs[1,2], Clara I. Sánchez[2], Stefan C. Saur[3], Thorsten Twellmann[3], Pim A. de Jong[4], and Bram van Ginneken[2,5]

[1] Fraunhofer MEVIS, Bremen, Germany
[2] Diagnostic Image Analysis Group, Radboud University Nijmegen Medical Centre, The Netherlands
[3] MeVis Medical Solutions AG, Bremen, Germany
[4] Department of Radiology, University Medical Center, Utrecht, The Netherlands
[5] Image Sciences Institute, University Medical Center, Utrecht, The Netherlands

Abstract. Ground glass nodules (GGNs) occur less frequent in computed tomography (CT) scans than solid nodules but have a much higher chance of being malignant. Accurate detection of these nodules is therefore highly important. A complete system for computer-aided detection of GGNs is presented consisting of initial segmentation steps, candidate detection, feature extraction and a two-stage classification process. A rich set of intensity, shape and context features is constructed to describe the appearance of GGN candidates. We apply a two-stage classification approach using a linear discriminant classifier and a GentleBoost classifier to efficiently classify candidate regions. The system is trained and independently tested on 140 scans that contained one or more GGNs from around 10,000 scans obtained in a lung cancer screening trial. The system shows a high sensitivity of 73% at only one false positive per scan.

Keywords: ground glass nodule, computer-aided detection, chest CT.

1 Introduction

Ground glass nodules (GGNs) are relatively rare findings in chest computed tomography (CT) examinations. These nodules have an increased attenuation but do not, like solid nodules, completely obscure the lung parenchyma [1], although they may have a solid component. It has been shown that GGNs have a much higher chance of being malignant than solid nodules [1]. Early detection of GGNs is therefore highly important. Beigelman-Aubry et al. [2] showed that both radiologists and computer-aided detection (CAD) systems designed for solid nodules have difficulties with detecting GGNs.

At present, no complete CAD system for GGNs has been tested on a large database. Kim et al. [3] described a slice-based CAD system using texture and intensity features that had a high false positive rate (FPR) and that was tested on only 14 patients. Zhou et al. [4] developed an automatic scheme for both

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 207–214, 2011.

detection and segmentation of GGNs based on vessel suppression, intensity and texture analysis. They reported high performance but the test data set contained only 10 GGNs. Ye et al. [5] presented a voxel-based method with rule-based filtering that was tested on 50 CT examinations with 52 GGNs. They reported a high sensitivity of 92.3% but also a high FPR of 12.7 per scan. Tao et al. [6] developed a multi-level detection scheme with classification at voxel-level and object-level. They focused on classification of small volumes of interest (VOIs) generated by a candidate detector algorithm which was not otherwise specified. The method was tested on a set of 1100 VOIs including 100 positive ones, from 153 healthy and 51 diseased patients. Results were provided for VOIs only, and neither the FPR per scan nor the total number of VOIs per scan were reported.

In this work, we focus on the automated detection of GGNs from thin-slice CT examinations. In contrast to the aforementioned works, the CAD system is tested on a large data set. A complete detection pipeline is presented, consisting of initial segmentation steps, candidate detection, feature extraction and classification. A comprehensive set of intensity and shape features are computed for each candidate region. As previous studies reported [3,7,8], false positive findings arise from partial volume effects of bronchovascular bundles, the chest wall, the dome of the diaphragm and large vessels. Therefore, we include context features that describe the position of the candidate region with respect to surrounding objects such as airways and the lung boundary.

2 Data and Experimental Design

Data for this study was provided by a large multi-center lung cancer screening trial with thin-slice, low-dose CT scans of current and former heavy smokers. We collected all CT examinations between April 2004 and April 2009 from one screening site, totaling around 10,000 scans from around 3,000 participants. From this data set, all scans in which at least one GGN was reported were selected. This resulted in 140 scans from 58 patients, including 76 unique GGNs. We considered GGNs in follow-up examinations as separate GGNs, leading to a total of 176 GGNs. For each GGN, a manual segmentation was provided. The effective diameter of the GGNs varied from 3.9 to 29.7 mm (median 13.9 mm). All CT examinations were performed with a slice thickness of 0.7 mm and the in-plane voxel size varied between 0.52 and 0.84 mm.

The data set was randomly split into two sets on a patient level, preventing data from the same patient being present in both sets. The training set consisted of 67 scans with 91 GGNs from 31 patients. The test set of 73 CT examinations with 85 GGNs from 27 patients was not touched during system development and was only used for evaluation of the final configuration of the CAD system.

3 Methods

Prior to candidate extraction, we apply a previously developed lung and airway segmentation algorithm [9,10] to each scan.

3.1 Initial Candidate Detection

The candidate detection procedure starts with applying a double-threshold density mask within the lung regions to obtain voxels with attenuation values defined as ground glass opacity. In this study, we used a range between -750 and -300 Hounsfield units (HU) [7,8]. At the boundaries of the lungs, vessels, and airways, partial volume effects lead to attenuation values in the defined range. Therefore, we apply a morphological opening operation using a spherical structuring element with a diameter of 3 voxels to remove the voxels at these edges from the density mask. After this, connected component analysis is performed to obtain candidate regions. Evidently, this results in a large amount of candidate regions. We eliminate all candidate regions which have a volume smaller than 34 mm^3 (volume of an ideal sphere with diameter of 4 mm). Current clinical guidelines [11] state that GGNs smaller than 5 mm do not require follow-up CT examinations and since volume measurements on CT have a certain variability due to partial volume effects, a safety margin of 1 mm is used in this system.

3.2 Features

We defined a rich set of features that can be subdivided into three categories:

Intensity features. The first group of intensity features consists of histogram statistics computed on a normalized histogram with a bin size of 1 HU. Four histograms are constructed from: voxels within the candidate mask, voxels within the bounding box defined by the candidate mask, voxels in the neighborhood created by dilating the candidate mask with a rectangular structuring element of size 3x3x3 voxels and similarly, but using a rectangular structuring element of size 5x5x5 voxels. The following histogram statistics are extracted: entropy, mean, height of mean bin, mode, height of mode bin and quantiles at 5%, 25%, 50%, 75% and 95%. Furthermore, we calculate the mean, standard deviation, minimum, maximum and the first 7 invariant Hu moments [12] over the intensity values of voxels within the candidate mask. Local binary patterns (LBP) [13] and Haar wavelets are used for texture analysis. Local binary patterns are computed from the bounding box defined by the candidate mask in which we resample this area to respectively a 16x16x16 and 32x32x32 volume. We apply the same histogram statistics to the histogram output of the LBP operator and use these as features. Using 2D Haar wavelets, all axial slices of the 32x32x32 resampled volume are decomposed into four bands. Then, all bands of the 32 slices are combined and histogram statistics are extracted from the horizontal, vertical and diagonal component of the combined high-frequency part. Finally, maximum vesselness [14] over multiple scales (1.0, 1.77, 3.16, 5.62, and 10.0 voxels) is computed for the voxels in the candidate mask and the mean and standard deviation of the vesselness values are used as features. The total number of intensity features is 103.

Shape features. Shape analysis of the candidate regions is performed using the binary mask of the candidate region. We calculate sphericity, compactness and

volume of the candidate region. In order to calculate the sphericity, we define a sphere S at the center of mass of the candidate region which has equal volume as the candidate region. Then, sphericity is defined as the ratio between the volume of the voxels of the candidate region within sphere S and the total volume of sphere S. For compactness, we used the ratio between the surface of the candidate region and its volume. Furthermore, the same set of 7 invariant Hu moments are computed from the candidate mask voxels to describe its appearance. Note that in contrast to the previous calculation of Hu moments, the voxels are in this case not weighed by their intensity value. This results in 10 shape features.

Context features. In the third category of features, the location of the candidate region in respect to the lung boundary and the airway tree is computed. For all voxels inside the candidate segmentation, the distance to the lung boundary and distance to the closest airway is computed. Then, the mean, standard deviation, minimum and maximum distance to the lung boundary and airways are computed and used as context features. Finally, using the lung segmentation, a bounding box is defined around the lungs. Using this bounding box, relative position features are computed, including relative X, Y and Z position, distance to center of mass of both lungs and distance to left bottom corner of bounding box. This yields 13 context features.

3.3 Classification

In the classification step, candidate regions are classified into GGN or non-GGN class using a two-stage classification approach. Note that in the second stage, the posterior probability of the first classifier is added as an additional feature.

In pilot experiments, we extensively tested different classifiers (Linear Discrimant Analysis (LDA), k-Nearest Neighbor (kNN) and GentleBoost [15]) for the first and second phase classification. In these experiments, 10-fold cross-validation on the training set was performed to test the performance of the different classifiers. Note that the 10 folds were again created by splitting the training set at a patient level. Consequently, all follow-up examinations of one patient were in the same fold to prevent bias.

For the first phase, we ranked all features according to Fisher's discriminant ratio [16] and we selected the four features with the highest ranking. Using these four features (two shape and two intensity features), LDA and kNN were tested and LDA proved to give slightly better results. Using the results from the 10-fold cross-validation on the training set, the posterior probability threshold for the first phase classification was determined. The threshold was set such that no true positives were lost in the first phase for the training set. This reduced the number of candidates in the training set by 66%. Consequently, all features only need to be calculated for about one third of all candidate regions, which accelerates the CAD system considerably ($\sim 40\%$).

For the second phase classification, we experimented with an LDA, kNN, and GentleBoost classifier. Optimal results for the kNN-classifier were found using k=60 and we used regression stumps as a weak classifier for the GentleBoost classification. Since the data set consists of a relatively high amount of features

Table 1. Performance of the CAD system for different feature groups. Sensitivity is reported at $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, 4, and 8 false positives per scan. The score is the average sensitivity at these 7 operating points.

Feature set	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	**1**	**2**	**4**	**8**	**Score**
All features	0.437	0.565	0.671	0.729	0.748	0.780	0.788	**0.674**
Shape and intensity features	0.476	0.569	0.650	0.686	0.705	0.765	0.772	**0.660**
Context and intensity features	0.465	0.578	0.645	0.687	0.710	0.749	0.767	**0.657**
Context and shape features	0.479	0.521	0.595	0.683	0.737	0.778	0.780	**0.654**
Intensity features	0.497	0.575	0.630	0.668	0.700	0.723	0.731	**0.647**
Context features	0.402	0.472	0.569	0.627	0.706	0.720	0.724	**0.603**
Shape features	0.486	0.506	0.550	0.604	0.637	0.708	0.724	**0.602**

(127) and a smaller amount of true positives (around 100), we decided to select the best 20 features using a Sequential Feed Forward Selection (SFFS) procedure to prevent overfitting of the classifier. During the SFFS procedure, the partial area under the curve (AUC) of the ROC curve was used as objective function. The upper threshold on false positive fraction was set at the value which corresponds to 5 false positives per scan. As the concept of boosting is based on sequentially applying weak classifiers on a subset of the data [15], feature selection was not used for the GentleBoost classifier. Finally, after extensive testing of different classifiers with combined feature selection, we concluded that the GentleBoost classifier had a slighty better performance and therefore we used this in the final configuration of the system.

In some cases multiple candidate regions were present for a single GGN. As we focused on detection, we counted a GGN as detected when at least one matching candidate was classified as positive. The remaining matching candidates were considered neutral in the evaluation and not counted as false negatives.

4 Results

The candidate detection step generated 524 $\pm$ 308 candidate regions per scan. Candidates are considered positive when the centers of mass of the GGN segmentation and the candidate region were within a distance d of each other. For $d = 10$ mm, the sensitivity of the candidate detector was 92% (84/91) and 87% (74/85) for the training and test set, respectively. For sake of readability, we omit a detailed reporting for a distance criterion of $d = 5$ mm where the results were comparable to the ones of $d = 10$ mm.

The first stage LDA classifier and second phase GentleBoost classifier were trained with all candidates from the training set and tested on the test set. After the first classification step, 32% of the candidate regions remained in the test set at the expense of eliminating three true positives. The FROC curve of the complete CAD system is given in Fig. 1 and sensitivities at various operating points are given in Table 1. At only one false positive per scan, 62 out of 85 GGNs (73% sensitivity) were detected.

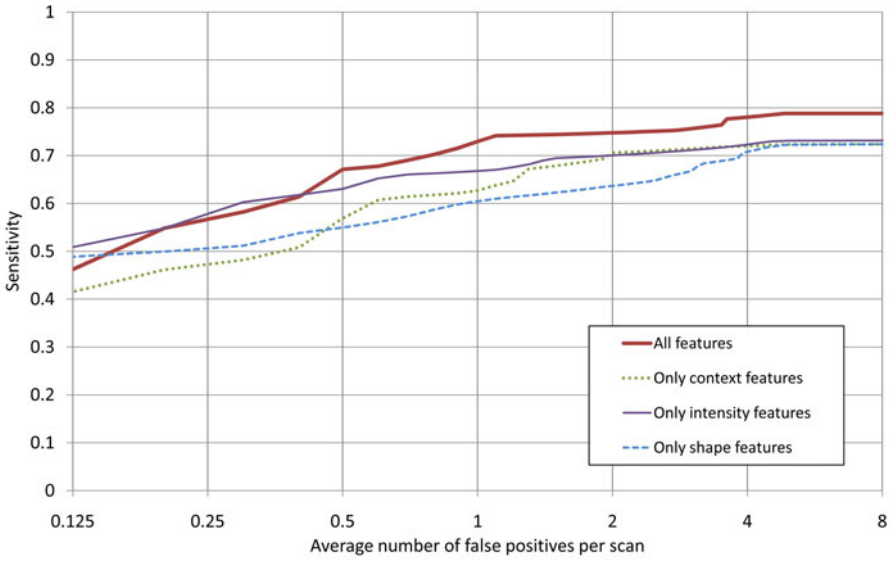

Fig. 1. FROC curves with a logarithmic x-axis. Results are shown for the proposed CAD system and systems that are only trained with one type of features.

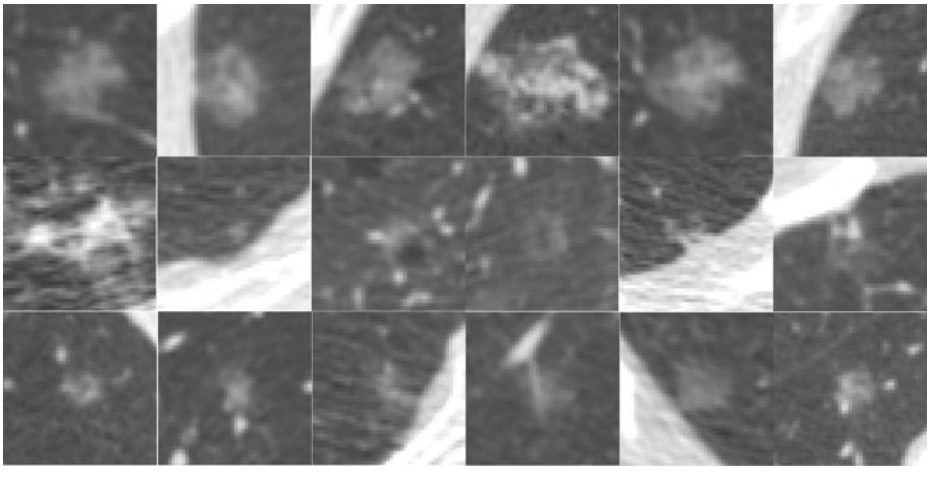

Fig. 2. Examples of true positives, false negatives and false positives of the CAD system. All images are axial views of 30 × 30 mm with a window level of -600/1600 HU. The top row shows the six true positives with the highest degree of suspicion in the test set according to the CAD system. The middle row shows six false negatives that the CAD system did not detect when set to operate at 1 false positive per scan. These nodules were picked up by the candidate detector, but they were not deemed suspicious enough. Finally, the bottom row shows the six false positives with the highest degree of suspicion in the test set according to the CAD system.

Furthermore, we investigated the effect of the three separate feature groups by using only one or only two in the second phase classification. Results are given in Fig. 1 and Table 1. An overall performance metric is derived from the FROC curves by averaging the sensitivities at $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, 4, and 8 false positives per scan. These results indicate that the combination of all features increases the performance considerably. In Fig. 2, we show examples of ground glass nodules which were correctly detected and missed by the CAD system.

5 Discussion and Conclusion

The FROC curve in Fig. 1 shows that our CAD system is able to find around 50% of all GGNs without any false positives. This is a very encouraging result. Moreover, a radiologist who retrospectively inspected the false positives of the system operating at 1 FP/scan (see Fig. 2, bottom row) indicated that many of these seemed to be GGNs. The reason why these findings had not been annotated may be that the scans in which these findings occurred contained multiple areas of ground glass opacity with a pattern resembling smoking related interstitial lung disease. On the other hand it may be that these were real GGNs missed by human readers. This interesting observation does require further study.

In clinical practice, a GGN CAD system will be used in combination with a solid nodule CAD system, which also produces false positives, and therefore we believe that the operating point at only 1 FP/scan is optimal. Moreover, the sensitivity does not increase much after 1 FP/scan.

From the 22 missed GGNs, 11 were actually missed in the candidate extraction step. We observed that in some cases a single GGN was detected as two separate candidate regions that were subsequently eliminated. Possible improvements are candidate clustering or integrating voxel classification in the candidate detection procedure, as done in [6]. Furthermore, even though the data for this study originated from over 10,000 scans obtained in a screening trial, the training set still contained less than 100 GGNs. We plan to collect a larger and more diverse training database in the future. This may help the CAD system to also recognize uncommon manifestations of GGNs, such as the ones on the second row of Fig. 2.

In conclusion, a complete computer-aided detection scheme for detection of ground glass nodules has been presented and tested on a large database. A comprehensive set of intensity, shape and context features was used to describe the appearance of a ground glass nodule. An optimized classification scheme using two stages of classification was employed. We evaluated the performance of the CAD system on an independent test set that was not touched during system development and obtained a sensitivity of 73% at only one false positive detection per scan. This is a substantially better performance than reported in previous work [3,6,4,5]. We are convinced that this performance level is sufficient for application of the system in clinical practice.

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. AJR Am. J. Roentgenol. 178(5), 1053–1057 (2002)
2. Beigelman-Aubry, C., Hill, C., Boulanger, X., Brun, A., Leclercq, D., Golmard, J., Grenier, P., Lucidarme, O.: Evaluation of a computer aided detection system for lung nodules with ground glass opacity component on multidetector-row CT. J. Radiol. 90(12), 1843–1849 (2009)
3. Kim, K.G., Goo, J.M., Kim, J.H., Lee, H.J., Min, B.G., Bae, K.T., Im, J.G.: Computer-aided diagnosis of localized ground-glass opacity in the lung at CT: Initial experience. Radiology 237, 657–661 (2005)
4. Zhou, J., Chang, S., Metaxas, D.N., Zhao, B., Ginsberg, M.S., Schwartz, L.H.: An automatic method for ground glass opacity nodule detection and segmentation from CT studies. In: IEEE EMBS, vol. 1, pp. 3062–3065 (2006)
5. Ye, X., Lin, X., Beddoe, G., Dehmeshki, J.: Efficient computer-aided detection of ground-glass opacity nodules in thoracic CT images. In: Proceedings of the 29th Annual International Conference of the IEEE EMBS. vol. 1, pp. 4449–4452 (2007)
6. Tao, Y., Lu, L., Dewan, M., Chen, A.Y., Corso, J., Xuan, J., Salganicoff, M., Krishnan, A.: Multi-level ground glass nodule detection and segmentation in CT lung images. In: Yang, G.Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 715–723. Springer, Heidelberg (2009)
7. Heitmann, K.R., Kauczor, H., Mildenberger, P., Uthmann, T., Perl, J., Thelen, M.: Automatic detection of ground glass opacities on lung HRCT using multiple neural networks. Eur. Radiol. 7(9), 1463–1472 (1997)
8. Kauczor, H.U., Heitmann, K., Heussel, C.P., Marwede, D., Uthmann, T., Thelen, M.: Automatic detection and quantification of ground-glass opacities on high-resolution CT using multiple neural networks: Comparison with a density mask. AJR Am. J. Roentgenol. 175(5), 1329–1334 (2000)
9. van Rikxoort, E.M., de Hoop, B., Viergever, M.A., Prokop, M., van Ginneken, B.: Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. Med. Phys. 36(7), 2934–2947 (2009)
10. van Ginneken, B., Baggerman, W., van Rikxoort, E.M.: Robust segmentation and anatomical labeling of the airway tree from thoracic CT scans. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 219–226. Springer, Heidelberg (2008)
11. Godoy, M.C.B., Naidich, D.P.: Subsolid pulmonary nodules and the spectrum of peripheral adenocarcinomas of the lung: recommended interim guidelines for assessment and management. Radiology 253(3), 606–622 (2009)
12. Hu, M.K.: Visual pattern recognition by moment invariants, computer methods in image analysis. IRE Transactions on Information Theory 8, 179–187 (1962)
13. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24(7), 971–987 (2002)
14. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: Wells, W., Colchester, A., Delp, S. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 130–137. Springer, Heidelberg (1998)
15. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. Ann. Stat. 28(2), 337–407 (2000)
16. Jobson, J.D.: Applied Multivariate Data Analysis. Springer, Heidelberg (1992)