Learning COPD sensitive filters in pulmonary CT

Lauge Sørensen[1], Pechin Lo[1], Haseem Ashraf[2], Jon Sporring[1], Mads Nielsen[1], and Marleen de Bruijne[1,3]

[1] Department of Computer Science, University of Copenhagen, Denmark
{lauges, pechin, madsn, sporring, marleen}@diku.dk
[2] Department of Respiratory Medicine, Gentofte University Hospital, Denmark
[3] Biomedical Imaging Group Rotterdam, Erasmus MC, the Netherlands

Abstract. The standard approaches to analyzing emphysema in computed tomography (CT) images are visual inspection and the relative area of voxels below a threshold (RA). The former approach is subjective and impractical in a large data set and the latter relies on a single threshold and independent voxel information, ignoring any spatial correlation in intensities. In recent years, supervised learning on texture features has been investigated as an alternative to these approaches, showing good results. However, supervised learning requires labeled samples, and these samples are often obtained via subjective and time consuming visual scoring done by human experts.

In this work, we investigate the possibility of applying supervised learning using texture measures on random CT samples where the labels are based on external, non-CT measures. We are not targeting emphysema directly, instead we focus on learning textural differences that discriminate subjects with chronic obstructive pulmonary disease (COPD) from healthy smokers, and it is expected that emphysema plays a major part in this. The proposed texture based approach achieves an 69% classification accuracy which is significantly better than RA's 55% accuracy.

1 Introduction

The traditional tools for diagnosis of chronic obstructive pulmonary disease (COPD) are pulmonary function tests (PFT)s. These are cheap and fast to acquire but suffer from several limitations, including insensitivity to early stages of COPD and lack of reproducibility [1]. More recently, computed tomography (CT) imaging has been used for direct measurement of one the components of the disease, namely emphysema, which is characterized by gradual loss of lung tissue and appears as low attenuation areas within the lung tissue. There are two common approaches for assessing emphysema in CT images: visual assessment, including sub-typing of emphysema based on radiological experience [2], and measures derived from the CT attenuation histogram, with the most widely used measure being relative area of voxels below a certain threshold (RA) [2].

RA disregards potentially valuable information in the CT image, such as spatial relations between voxels. Various alternatives have been suggested for analyzing emphysema in CT images. One such approach is analysis of bullae size distribution [3]. Another approach is supervised texture classification where a classifier is trained on manually annotated regions of interest (ROI)s [4–7]. The output of a trained classifier can be used for COPD quantification by fusion of individual ROI posterior probabilities [7].

Supervised learning requires a training set with labeled data which is usually acquired by manual annotation. However, having human observers manually annotating ROIs can be problematic. First of all, it is a subjective process suffering from inter-observer variability. This problem can partly be addressed by consensus readings of several experts. Another drawback is the time needed for doing the annotations, and when the data set is large, manual annotation is infeasible. Further, analysis will be limited to current knowledge and experience of experts, and there can be a bias towards typical cases in the annotated data set. In the emphysema case, this means restricting ourselves to the three known radiographic subtypes of emphysema [2].

In this work, we explore the possibility of diagnosing COPD in volumetric CT images of the lung based on texture classification without manual labeling. PFTs are used to define two subject groups, a healthy and a COPD group, and ROIs are randomly sampled from these two groups and labeled according to group membership. A supervised learning framework is applied for learning filters that are able to separate the two groups. This approach is less committed, objective, and can potentially uncover new textural patterns, or emphysema subtypes, as being part of COPD.

Classification is based on the k nearest neighbor (kNN) classifier using dissimilarity between sets of feature histograms as distance, and the features are based on a rotation invariant, multi-scale Gaussian filter bank [8]. The classification framework used here is similar to the one used in [6], but with a larger set of filters and in 3D instead of 2D. The obtained results are compared to RA in the experiments.

2 Selection of training samples

The classification framework relies on a grouping of the CT images into different subject groups, according to non-CT measures, and a lung segmentation S obtained from each CT image I. ROIs are sampled at random within the lung fields in the images and assigned labels, ω_i, according to subject group membership. The lung segmentation S is used for two purposes. First of all, it is used for limiting the random sampling to the lung fields. Secondly, it is used for allowing only lung parenchyma voxels to contribute to the obtained feature histograms as described in Section 3. In this work, we use PFTs to group the CT images, and S is extracted from I using thresholding and morphological smoothing, similar to [9].

3 Texture measures

Each ROI is represented by a set of feature histograms representing distributions of filter responses computed in the ROI. The filtering is done by normalized convolution [10] with a binary mask to exclude contribution from larger non-parenchyma structures, such as the trachea, the main bronchi, and the exterior of the lung. A rotation invariant, multi-scale Gaussian filter bank [8] comprising eight basis filters is used.

3.1 Filters

Eight different measures of local image structure are used as base filters: the Gaussian function G_σ; the three eigenvalues of the Hessian $\lambda_{i,\sigma}, i = 1, 2, 3$, ordered such that $|\lambda_{1,\sigma}| \geq |\lambda_{2,\sigma}| \geq |\lambda_{3,\sigma}|$; gradient magnitude $||\nabla G_\sigma||_2 = \sqrt{I_{x,\sigma}^2 + I_{y,\sigma}^2 + I_{z,\sigma}^2}$, where $I_{x,\sigma}$ denotes the partial first order derivative of image I w.r.t. x at scale σ; Laplacian of the Gaussian $\nabla^2 G_\sigma = \lambda_{1,\sigma} + \lambda_{2,\sigma} + \lambda_{3,\sigma}$; Gaussian curvature $K_\sigma = \lambda_{1,\sigma}\lambda_{2,\sigma}\lambda_{3,\sigma}$; and the Frobenius norm of the Hessian $||H_\sigma||_F = \sqrt{\lambda_{1,\sigma}^2 + \lambda_{2,\sigma}^2 + \lambda_{3,\sigma}^2}$.

The filtering is performed by normalized convolution [10] with a Gaussian function

$$I_\sigma = \frac{(S(\mathbf{x})I(\mathbf{x})) * G_\sigma(\mathbf{x})}{S(\mathbf{x}) * G_\sigma(\mathbf{x})},$$

where $*$ denotes convolution and segmentation S computed from image I is used as an indicator function, indication whether voxel $\mathbf{x} = [x, y, z]^T$ is a lung parenchyma voxel or not. Derivatives are computed on the filtered images using finite differences.

3.2 Histogram estimation

The filter responses are quantized into feature histograms. The bins edges are derived using adaptive binning [11]. This technique locally adapts the histogram bin widths to the data set at hand such that each bin contains the same mass when computing the histogram of all data while disregarding class labels. Only voxels within a lung segmentation S are used, and the resulting histogram is normalized to sum to one.

4 Classification

Classification is performed using the kNN classifier with summed histogram dissimilarity as distance

$$D(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N_f} L(f_i(\mathbf{x}), f_i(\mathbf{y})), \tag{1}$$

where N_f is the number of feature histograms, $L(\cdot,\cdot)$ is a histogram dissimilarity measure, and $f_i(\mathbf{x}) \in \mathbb{R}^{N_b}$ is the i'th feature histogram with N_b bins estimated from an ROI centered on $\mathbf{x}$. Two histogram dissimilarity measures L are considered: L1-norm and L2-norm. The L1-norm and L2-norm are instances of the p-norm

$$L_p(H,K) = ||H-K||_p = \left(\sum_{i=1}^{N_b} |H_i - K_i|^p \right)^{1/p} ,$$

with $p = 1$ or $p = 2$ and where $H \in \mathbb{R}^{N_b}$ and $K \in \mathbb{R}^{N_b}$ are histograms each with N_b bins.

The posterior probability of belonging to class ω_i given that the current ROI is centered on voxel $\mathbf{x}$ is estimated in the kNN classifier by $P(\omega_i|\mathbf{x}) = k_{\omega_i}(\mathbf{x})/k$, where $k_{\omega_i}(\mathbf{x})$ is the number of nearest neighbors according to (1) belonging to class ω_i obtained from a total of k nearest neighbors.

5 Experiments

5.1 Data

Experiments are conducted using 296 low-dose volumetric CT images (tube voltage 140 kV, exposure 40 mAs, slice thickness 1 mm, and in-plane resolution ranging from 0.72 to 0.78 mm) from 296 different (ex-)smokers enrolled in the Danish Lung Cancer Screening Trial [12].

Two subjects groups, $\omega_i = \{\text{healthy}, \text{COPD}\}$, are defined based on the GOLD criteria [13]. These criteria use two PFTs based measures: expiratory volume in one second over forced vital capacity (FEV$_1$/FVC), and forced expiratory volume in one second corrected for age, sex, and height (FEV$_1$%pred). The healthy group is defined by FEV$_1$%pred ≥ 80 and FEV$_1$/FVC ≥ 0.7. The COPD group is defined by FEV$_1$%pred $< 80\%$ and FEV$_1$/FVC < 0.7, which corresponds to GOLD stage II or higher [13]. The healthy group contains 144 CT images and the COPD group contains 152 CT images. For each CT image, 50 cubic $r \times r \times r$ ROIs are sampled at random, thus a total of 14800 ROIs are used in the experiments. A separate set of 10 ROIs per subject is sampled to compute the adaptive binning described in Section 3.2.

Since PFTs are not very reproducible [1], the grouping is enhanced by ensuring that the criteria are fulfilled at two time instances; both when the CT images were acquired and one year after.

5.2 Training and parameter selection

There are several parameters to set in the classification system: ROI size r, number of histogram bins N_b, k in the kNN classifier, histogram dissimilarity measure L, and which filters, out of the ones described in Section 3.1, to use at which scales σ. In this work, we use $N_b = \sqrt[3]{\text{number of voxels in the ROI}} = r$ bins. This ensures that the standard deviation across bins is proportional to the standard deviation within bins. The best combination of $r = \{21, 31, 41\}$, $L = \{L_1, L_2\}$, and $k = \{25, 35, 45\}$ is learned using cross-validation, and sequential forward feature selection (SFS) is used for determining the optimal filter subset for each combination. The scales of the filters are sampled exponentially according to $\sigma_i = 0.6(\sqrt{2})^i$ mm, $i = 0, \ldots, 6$. Together with the original intensity, this amounts a total of 57 feature histograms considered in the feature selection.

The CT images in the training set are divided into two sets by randomly placing half the images of each group in each set. The classification system is trained, and parameters are tuned by using one set as prototypes in the kNN classifier and by choosing the features and parameter settings that minimize the classification error on the other set.

5.3 Evaluation

The performance is estimated using 3-fold cross-validation, training the classifier as described above and applying the best performing kNN classifier, in terms of validation error, with the training set as prototypes to the test set. The results are evaluated in three ways. First, by maximum a posteriori classification accuracy on the ROIs. For the remaining cases, each subject is measured by posterior fusion: the mean healthy posterior probability is computed across all sampled ROIs in the subject. The second evaluation is maximum a posteriori classification accuracy on subject level, and the third is the ability to separate the healthy group of subjects from the COPD group according to a rank sum test on the mean healthy posterior.

Since the texture based CT measurements are proposed as an alternative to RA, we compare the obtained results to RA, computed both on the sampled ROIs and on whole lungs. The best RA for ROI classification, RA_i, is determined using cross-validation on the same data sets as used when training the kNN classifier on the range $i = [-960, -950, \ldots, -890]$ HU. Thresholds in this range are commonly used when measuring emphysema in CT [2]. The best percentage threshold used for classification based on RA is also determined during this procedure.

5.4 Results

The filters selected by SFS using the best parameter setting are shown in Table 1. Three of the filters are selected in two out of three folds. The selected optimal kNN parameters are $r = 41$ in all three folds and ordered by fold $L = L_2, L_1, L_2$ and $k = 35, 45, 35$. The selected optimal RA parameters are, ordered by fold, HU threshold $= -890, -960, -890$ and percentage threshold $= 32, 15, 47$. ROI classification accuracies, subject classification accuracies, areas under the receiver operating characteristic curve (AUC), and p-values for difference between groups according to a rank sum test are shown in Table 2. RA using -950 HU is also included for the sake of completeness. kNN achieves significantly higher ROI and subject classification accuracy than RA, $p < 10^{-4}$ according to McNemar's test. RA is able the pick up an overall group effect but has poor discrimination

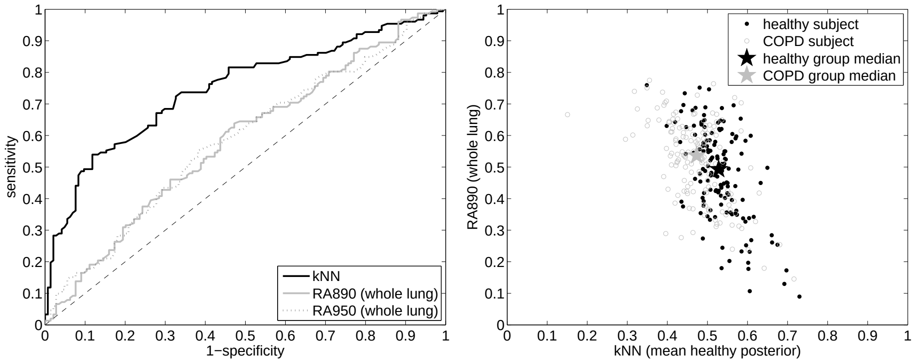
Table 1. Selected filters in kNN in the cross-validation procedure.

Fold	Selected filters
1	$\nabla^2 G_{0.6}, \lambda_{2,0.6}, \lambda_{3,0.6}, \lambda_{2,0.85}, \|\|\nabla G_{2.4}\|\|_2, \lambda_{1,2.4}, \nabla^2 G_{3.4}, K_{3.4}, \|\|\nabla G_{4.8}\|\|_2$
2	$K_{0.6}, G_{1.2}, \|\|\nabla G_{1.2}\|\|_2, \|\|\nabla G_{4.8}\|\|_2, K_{4.8}$
3	$\nabla^2 G_{0.6}, \|\|H_{0.6}\|\|_F, \|\|\nabla G_{1.7}\|\|_2, \lambda_{1,1.7}, G_{2.4}, K_{4.8}$

Table 2. Classification accuracies, AUCs, and p-values from a rank sum test.

Measure	ROI accuracy	Subject accuracy	AUC	p-value
kNN	0.58	0.69	0.75	$< 10^{-4}$
RA$_{\text{learned}}$ (ROIs)	0.53	0.55	-	-
RA$_{890}$ (whole lung)	-	-	0.58	0.012
RA$_{950}$ (whole lung)	-	-	0.59	0.012

ability on a subject level. Figure 1(a) shows receiver operating characteristic (ROC) curves for the kNN classifier on the ROIs as well as for RA on the whole lung field. In the case of kNN, the parameter being varied is the healthy posterior threshold, and in the case of RA, the parameter is the threshold on the percentage of low attenuation voxels. AUC is clearly larger for the kNN classifier compared to RA. Figure 1(b) shows the COPD measures obtained by mean kNN healthy posterior and RA$_{890}$. The separation between the groups is much better for kNN as indicated by the classification accuracies and p-values in Table 2.



(a) (b)

Fig. 1. (a) ROC curves at subject level. The curve for kNN is based on mean healthy posterior computed for each subject. (b) Scatter plot of mean kNN healthy posterior versus RA$_{890}$.

6 Discussion

The proposed texture based method outperforms RA in all comparisons. The classification accuracies in Table 2 are significantly higher, both at ROI and at subject level. When computing RA on the full lung, which is the common way of applying RA [2], more information is used than is available to the kNN classifier, and even in these cases RA performs worse.

The ROI classification accuracies in Table 2 are relatively low when compared to accuracies reported in the literature [4–7]. However, it is important to note that in the cited cases, the ROIs and labels are obtained by manual labeling of "interesting" areas in CT images. In this work, no manual labeling has been done, instead the labels were obtained by taking random samples of lung tissue within the lung fields. We expect the COPD group to also contain samples with no apparent lung disease pattern in a random sampling setup, hence, the classes are likely to overlap more when using this approach.

Intensity can be directly related to emphysema since emphysematous regions have lower intensities due to loss of lung tissue, and therefore original and smoothed intensities are considered important features. Nevertheless, intensity is not selected in the first cross-validation fold, but the Laplacian of the Gaussian which approximates the zero-order information at a larger scale is selected and may compensate for this. Three filters are selected in two out of three cross-validation folds: Laplacian of the Gaussian, $\nabla^2 G_{0.6}$, gradient magnitude, $||G_{4.8}||_2$, and Gaussian curvature $K_{4.8}$. $\nabla^2 G_{0.6}$ can be seen as a blob detector at low scale, and it may be small low attenuation areas within the lung tissue that are picked up by this filter. $||G_{4.8}||_2$ measures large scales edges and $K_{4.8}$ measures large scale blobs.

Emphysema is not uniformly distributed within the lungs. Paraseptal emphysema is located in the periphery of the lung, and centrilobular emphysema is predominantly in the upper lobes [2]. It would therefore be interesting to see whether the COPD related textural differences found in this work are localized in specific regions of the lungs.

Subjects were grouped using PFTs, and thus the classification system is trained to imitate diagnosis of COPD based on PFT. As can be seen from Figure 1(b) and the reported numbers, the learned filters achieve this to some degree. The result is a quantitative measure of COPD which may be more sensitive and reproducible than PFT. This facilitates study of disease development and progression in large cohorts such as current lung cancer screening trials, which may help improve the understanding of pathogenesis of COPD and eventually lead to improved diagnosis, prognosis, and treatment of individuals.

In summary, we conclude that it is possible to learn COPD sensitive filters in CT in a less committed, data-driven manner without, the often tedious, manual annotation of data. A kNN classifier using texture measures based on these filters is capable of separating healthy subjects from subjects with COPD, when these are diagnosed based solely on PFTs.

Acknowledgements. This work is partly funded by the Danish Council for Strategic Research (NABIIT), the Netherlands Organisation for Scientific Research (NWO), and AstraZeneca, Lund, Sweden.

References

1. Dirksen, A., Holstein-Rathlou, N.H., Madsen, F., Skovgaard, L.T., Ulrik, C.S., Heckscher, T., Kok-Jensen, A.: Long-range correlations of serial FEV1 measurements in emphysematous patients and normal subjects. J Appl Physiol **85**(1) (Jul 1998) 259–265
2. Webb, W.R., Müller, N., Naidich, D.: High-Resolution CT of the Lung, Third Edition. Lippincott Williams & Wilkins (2001)
3. Blechschmidt, R.A., Werthschützky, R., Lörcher, U.: Automated CT image evaluation of the lung: a morphology-based concept. IEEE Trans Med Imaging **20**(5) (May 2001) 434–442
4. Mendonça, P.R.S., Padfield, D.R., Ross, J.C., Miller, J.V., Dutta, S., Gautham, S.M.: Quantification of emphysema severity by histogram analysis of CT scans. In: MICCAI 2005. Volume 3749 of LNCS. (Sep 2005) 738–744
5. Xu, Y., Sonka, M., McLennan, G., Guo, J., Hoffman, E.A.: MDCT-based 3-D texture classification of emphysema and early smoking related lung pathologies. IEEE Trans Med Imaging **25**(4) (Apr 2006) 464–475
6. Sørensen, L., Shaker, S.B., de Bruijne, M.: Texture classification in lung CT using local binary patterns. In: MICCAI 2008. Volume 5241 of LNCS. (Sep 2008) 934–941
7. Park, Y.S., Seo, J.B., Kim, N., Chae, E.J., Oh, Y.M., Lee, S.D., Lee, Y., Kang, S.H.: Texture-based quantification of pulmonary emphysema on high-resolution computed tomography: Comparison with density-based quantification and correlation with pulmonary function test. Investigative Radiology **43**(6) (Jun 2008) 395–402
8. ter Haar Romeny, B.M.: Applications of scale-space theory. In: Gaussian Scale-Space Theory. Dordrecht: Kluwer Academic Publishers (1997) 3–19
9. Hu, S., Hoffman, E., Reinhardt, J.: Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. IEEE Trans Med Imaging **20**(6) (Jun 2001) 490–498
10. Knutsson, H., Westin, C.F.: Normalized and differential convolution: Methods for interpolation and filtering of incomplete and uncertain data. In: CVPR. (Jun 1993) 515–523
11. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1) (Jan 1996) 51–59
12. Pedersen, J.H., Ashraf, H., Dirksen, A., Bach, K., Hansen, H., Toennesen, P., Thorsen, H., Brodersen, J., Skov, B.G., Døssing, M., Mortensen, J., Richter, K., Clementsen, P., Seersholm, N.: The Danish randomized lung cancer CT screening trial–overall design and results of the prevalence round. J Thorac Oncol **4**(5) (May 2009) 608–614
13. Rabe, K.F., Hurd, S., Anzueto, A., Barnes, P.J., Buist, S.A., Calverley, P., Fukuchi, Y., Jenkins, C., Rodriguez-Roisin, R., van Weel, C., Zielinski, J.: Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. Am J Respir Crit Care Med **176**(6) (Sep 2007) 532–555