Using a priori knowledge to classify in vivo images of the lung

Chesner Désir, Caroline Petitjean, Laurent Heutte, Luc Thiberville

Using a priori knowledge to classify in vivo images of the lung

Chesner Désir[1], Caroline Petitjean[1], Laurent Heutte[1], and Luc Thiberville[2]

[1] Université de Rouen, LITIS EA 4108
BP 12, 76801 Saint-Etienne-du-Rouvray, France,
[2] CHU de Rouen, LITIS EA 4108, 76031 Rouen, France
{chesner.desir,caroline.petitjean,laurent.heutte,luc.thiberville}@univ-rouen.fr

Abstract. Until recently, the alveolar region could not be investigated in-vivo. A novel technique, based on confocal microscopy, can now provide new images of the respiratory alveolar system, for which quantitative analysis tools must be developed, for diagnosis and follow up of pathological situations. In particular, we wish to aid the clinician by developing a computer-aided diagnosis system, able to discriminate between healthy and pathological subjects. This paper describes this system, in which images are first characterized through a 148-feature vector then classified by an SVM (Support Vector Machine). Experiments conducted on smoker and non smoker images show that the dimensionality of the feature vector can be reduced significantly without decreasing classification accuracy, and thus gaining some insight about the usefulness of features for medical diagnosis. These promising results allow us to consider interesting perspectives for this very challenging medical application.

1 Introduction

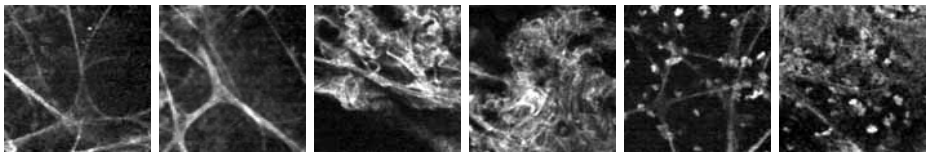
The lungs are the essential respiration organ. They are divided into two anatomic and functional regions: (i) the air conduction system, that includes the trachea, bronchi, and bronchioles, and (ii) the gas-exchange region, or lung parenchyma, made of alveolar sacs. These sacs are made up of clusters of alveoli, tightly wrapped in blood vessels, that allow for gas exchange. Whereas the conduction airways can be explored in vivo during bronchoscopy, the alveolar region was until recently unreachable for in vivo morphological investigation. Therefore, the pathology of the distal lung is currently assessed only in vitro, using invasive techniques such as open lung biopsies. No real time imaging was available.

Recently, a new endoscopic technique, called Fibered Confocal Fluorescence Microscopy (FCFM), has been developed that enables the visualisation of the more distal regions of the lungs in-vivo [1]. The technique is based on the principle of fluorescence confocal microscopy, where the microscope objective is replaced by a fiberoptic miniprobe, made of thousands of fiber cores. The miniprobe can be introduced into the 2 mm working channel of a flexible bronchoscope to produce in-vivo endomicroscopic imaging of the human respiratory tract in real time. Real-time alveolar images are continuously recorded during

the procedure and stored for further analysis. This very promising technique could replace lung biopsy in the future and might prove to be helpful in a large variety of diseases, including interstitial lung diseases [2].

In this context, a clinical trial is currently being conducted that collects FCFM images in several pathological conditions of the distal lungs and on healthy smoker and non-smoker volunteers. FCFM images represent the alveolar structure, made of elastin fiber (Figure 1), with an approximate resolution of 1μm per pixel. This structure appears as a network of (almost) continuous lines. This elastic fiber framework can be altered by distal lung pathologies and as one can see on Figure 1, images acquired on pathological subjects differ from the ones acquired on healthy subjects.

We describe in this paper a first attempt to classify FCFM images as healthy or pathological. We have designed a 148-feature vector to describe the images, the discriminative power of which is assessed through a leave-one-out evaluation of a 1-Nearest Neighbour (1-NN) classifier. We then show that the size of this feature vector can be reduced significantly without decreasing classification accuracy by using an SVM wrapper-based feature selection technique. We thus show how to gain some insight about usefulness of the features for the discrimination of healthy/pathological FCFM images. The remaining of this paper is organized as follows: our classification method is described in Section 2, and results and discussion are provided in Section 3. Section 4 concludes and draws some perspectives for this work.



(a) NS healthy (b) NS healthy (c) NS patho. (d) NS patho. (e) S healthy (f) S patho.

Fig. 1. FCFM images in non-smoking (NS) and smoking (S) subjects. In smoker images, the white spots are macrophages, which are cells normally invisible in non-smoker but made visible by the smoke trapped in it.

2 Feature extraction and classification

2.1 Feature extraction

Features must be chosen to allow the discrimination between healthy and pathological subjects. Despite the novelty of the images, their visual analysis allows to highlight some differences, that can be used as a priori knowledge to design the feature vector. The alveolar structure in healthy subjects can be described

as contrasted continuous lines and curves. On the opposite, in the pathological subset, the disorganization of the meshing is illustrated by the numerous irregularities and the tangle of the fibered structures (Figure 1). Differences are mostly visible for the structure shape, image texture and contrast.

The **structure contrast** can be characterized by studying (i) first order statistics on the image histogram: mean, variance, skewness, kurtosis, entropy, (ii) pixel densities obtained on binarized images using Otsu thresholding, and (iii) the sum of the image gradient values, obtained using Prewitt operator. We could suppose that pathological images will have higher values of densities than healthy ones because of an emphasized disorganization of the meshing in pathological images.

The **complexity of the structure shape** can be characterized by studying the image skeleton. After skeletonization [3] obtained on the binary image, the number of junction points is computed. One can suppose that on clearly organized, healthy images, this number will be small, contrary to pathological images where the meshing mess will induce a higher number of points.

The **image texture** can be characterized by Haralick parameters computed from co-occurrence matrix [4]. Co-occurrence matrix provides the joint distribution of gray-level intensities between two image points. These two points are located according several configurations, that represent different distances and rotation angles. We chose the following classical 10 translation vectors: [0 1], [-1 1], [-1 0], [-1 -1], [0 2], [-1 2], [-1 -2], [-2 1], [-2 0], [-2 -1]. From the features originally proposed by Haralick, we retain the following ones: energy, contrast, correlation, variance, inverse different moment, entropy, sum average, sum entropy, sum variance, difference entropy, difference variance, and two information measures of correlation. The only discarded feature is the maximum correlation coefficient, which is too computationally expensive. To these 13 parameters we added dissimilarity, a measure of homogeneity [5]. All these 14 parameters are computed over the 10 co-occurrence matrices (Table 1).

Table 1. Features used to characterize FCFM images

	Features	Number
Contrast	Histogram statistics	5
	Pixel density	1
	Sum of image gradient	1
Shape	Number of junction points in skeleton	1
Texture	Haralick parameters	140
	Total	148

2.2 Classifier

On the previously cited features several state-of-the-art classifiers have been implemented. First a 1-Nearest Neighbour (1-NN) classifier is used to assess the discriminating power of the features. Due to the high computational cost of the 1-NN classifier, we have also implemented a Support Vector Machine (SVM) classifier on our features [6]. The SVM classifier, one of the most performing and most used classification algorithm, is a binary classifier algorithm that looks for an optimal hyperplane as a decision function in a high-dimensional space. A classical choice for the kernel is the cubic polynomial kernel.

In order to improve the prediction performance of the classifier, and to provide faster and more cost-effective decision, variable selection [7] can be used. It can also provide a better understanding of which visual features discriminate the data. Support Vector Machine -Recursive Feature Elimination (SVM-RFE) is one way to perform variable selection [8]. The goal is to find a subset of size r among d variables ($r < d$) which maximizes the performance of the predictor. The method is based on a sequential backward selection. One feature at a time is removed until r features are left. The removed variables are the ones that minimize the variation of the margin.

2.3 Experimental protocol

Because of the relatively small number of images in the non-smoker and the smoker bases, a leave-one-out cross validation process is used, which ensures unbiased generalization error estimation. It consists in extracting one sample image from the image base for validation, the rest of the base being used for learning. Recognition rate is computed over all the samples.

Table 2. Number of images in the non-smoker and smoker databases

	Non-smoker database	Smoker database
Healthy subjects	35	58
Pathological subjects	43	39
Total	78	97

3 Results

The SVM classifier and SVM-RFE based feature selection [8] are implemented using the SVM and Kernel Methods Matlab Toolbox [9]. The system performance is assessed with correct classification rate, error rate, false negative rate (FN), which is the proportion of healthy instances that were erroneously reported as pathological and false positive rate (FP), which is the proportion of

pathological cases considered healthy.

Results obtained with the 1-NN classifier are shown in Table 3. Let us recall that the 1-NN classifier is used here to assess the discriminative power of the feature set. As one can see in Table 3, the feature set seems to be better adapted to the discrimination of healthy/pathological non-smoker images. This can be explained by the presence of macrophages and smoke trapped in the alveolar walls in smoker images. Indeed, the line network is hidden behind the macrophages, making it difficult to characterize the structure. On the other hand, recognition rates of 95% and 89% indicate that room for improvement is left with this feature set.

Table 3. Results provided by 1-NN classifier

	Non-smoker database	Smoker database
Recog. rate	95%	89%
Error rate	5%	11%
FN	6%	9%
FP	5%	15%

Results obtained with the SVM and SVM-RFE are shown in Table 4. They are quite satisfying for the considered databases. Thanks to feature selection, the number of features, initially 148, drops down to 20 for non-smoker images, and 36 for smoker images, without decreasing classification performance. The selection of relevant variables allows to gain some insight about the usefulness of features: the most discriminating ones for non smoker images are the number of junction points, the contrast, the difference variance, and correlation computed from co-occurrence matrices, which highlights the importance of local, contrast-based differences between healthy and pathological subjects. On the other hand, for smoker images, retained features include the sum of image gradient, the sum variance, variance and contrast. Note that the number of junction points is no more considered for smoker images, which can be explained by the fact that the line network is hidden behind the macrophages. Finally, the reduced feature sets obtained on smoker and non-smoker images can be compared: only 6 features are jointly retained, which confirms that the discrimination of healthy/pathological images should be investigated separately for smokers and non-smokers.

4 Conclusions

The present work deals with the classification of a new category of images from the distal lung. The images were acquired using a fibered confocal fluorescence microscopy, a technique that enables the observation of in vivo alveolar structures for the first time. Such images are not well described so far, and difficult to

Table 4. Results provided by SVM and SVM-RFE classifier

	Non-smoker database		Smoker database	
	SVM	SVM-RFE	SVM	SVM-RFE
Feature number	148	20	148	36
Recog. rate	92%	97%	94%	94%
Error rate	8%	3%	6%	6%
FN	9%	3%	5%	5%
FP	7%	2%	8%	7%

discriminate by pathologists and respiratory physicians. Our classification system, that aims at discriminating healthy cases from pathological ones, shows satisfying performance for non-smoker and smoker images. However, the corresponding database should be extended to confirm these results. Because the clinical trial is ongoing, this will be feasible in the near future. Results could still be improved by using other texture-oriented features such as the local binary patterns, as well as more reliable classifiers such as random forests for example.

Future work will also concern rendering the process real-time, so as to aid the clinician during in vivo examination. Classification methods could also give information about which part of the image is the most discriminant or which part of the structure might be more altered by pathologies. A future goal will also be to discriminate between different pathologies: interstitial lung diseases (such as systemic sclerosis, fibrosis, sarcoidosis), carcinomatous lesions etc.

References

1. Thiberville, L., Moreno-Swirc, S., Vercauteren, T., Peltier, E., Cave, C., Bourg-Heckly, G.: In vivo imaging of the bronchial wall microstructure using fibered confocal fluorescence microscopy. American Journal of Respiratory and Critical Care Medicine **175** (October 2007) pp. 178–187
2. Thiberville, L., G.Bourg-Heckly, M. Salaün, S.D., Moreno-Swirc, S.: Human in-vivo confocal microscopic imaging of the distal bronchioles and alveoli. Chest Journal **132**(4) (2007) 426
3. Dibajaa, G.S., Thiel, E.: Skeletonization algorithm running on path-based distance maps. Image and Vision Computing **14:47-57** (1996)
4. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. Systems, Man and Cybernetics **3(6)** (1973)
5. Pratt, W.: Digital Image Processing, 2nd Edition. John Wiley & Sons (1991)
6. Vapnik, V.: The nature of statistical learning theory. N-Y: Springer-Verlag (1995)
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3** (2003) 1157–1182
8. Rakotomamonjy, A.: Variable selection using SVM-based criteria. Journal of Machine Learning Reseach (2003) 3:1357–1370
9. Canu, S., Grandvalet, Y., Guigue, V., Rakotomamonjy, A.: SVM and kernel methods matlab toolbox. Perception Systemes et Information, INSA de Rouen, France (2005)