



---

# **Quantitative Analysis for Pneumoconiosis Screening and Staging on Digital Chest Radiographs**

**Hao Xu, Weizhong Yan, Ramasubramanian Sundararajan,  
Xiaodong Tao and Chao Yang**

**2010GRC524, May 2010**

**Public (Class 1)**

Technical Information Series

Copyright © 2010. General Electric Company. All rights reserved.

## GE Global Research

### Technical Report Abstract Page

**Title** Quantitative Analysis for Pneumoconiosis Screening and Staging on Digital Chest Radiographs

<b>Author(s)</b>	Hao Xu	<b>Phone</b>	86 21 3877-2186
	Weizhong Yan		1 518-387-5704
	Ramasubramanian Sundararajan		91 80-4088-1325
	Xiaodong Tao		1 518-387-7502
	Chao Yang*		-

**Component** Imaging Technologies Laboratory, ShangHai

<b>Report Number</b>	2010GRC524	<b>Date</b>	May 2010
----------------------	------------	-------------	----------

<b>Number of Pages</b>	08	<b>Class</b>	Public (Class 1)
------------------------	----	--------------	------------------

**Key Words:** computer aided detection (CAD), active shape model (ASM), support vector machine (SVM), pneumoconiosis, digital radiography.

**Abstract:** Quantitative Analysis for Pneumoconiosis Screening and Staging on Digital Chest Radiographs Abstract: This paper presents a computer aided detection scheme on digital chest radiographs for pneumoconiosis screening and staging. The scheme involves several medical image processing and analysis technologies, i.e. lung segmentation algorithm on digital chest X-rays by using the active shape model, image enhancement and features extraction from lung regions in the preprocessed image, feature down-selection by correlation analysis and clustering method, classification on lung regions by multi-class classifier, support vector machine, and finally, the quantitative report and staging result based on the chest generated by the classification probabilities and the staging procedure from diagnostic criteria. Experiments are conducted on our digital chest x-ray database. Primary result shows the classification performance on the test set is sensitivity 98.1%, specificity 79.6%, and accuracy 81.8%.

Manuscript received May 25, 2010

# Quantitative Analysis for Pneumoconiosis Screening and Staging on Digital Chest Radiographs

**Abstract:** This paper presents a computer aided detection scheme on digital chest radiographs for pneumoconiosis screening and staging. The scheme involves several medical image processing and analysis technologies, i.e. lung segmentation algorithm on digital chest X-rays by using the active shape model, image enhancement and features extraction from lung regions in the pre-processed image, feature down-selection by correlation analysis and clustering method, classification on lung regions by multi-class classifier, support vector machine, and finally, the quantitative report and staging result based on the chest generated by the classification probabilities and the staging procedure from diagnostic criteria. Experiments are conducted on our digital chest x-ray database. Primary result shows the classification performance on the test set is sensitivity 98.1%, specificity 79.6%, and accuracy 81.8%.

**Keywords:** Computer aided detection (CAD), active shape model (ASM), support vector machine (SVM), pneumoconiosis, digital radiography

## 1 Introduction

Pneumoconiosis is a lung disease caused by the long-term inhalation of operative dust, such as coal, asbestos and silica, and the local tissue reaction to the accumulated dust particles [1]. Chest radiography is the most practical tool according to the guidelines of China and International Labor Organizer (ILO) standard for pneumoconiosis diagnosis, which is based on the assessment of opacities on chest X-Ray (CXR) films and comparison with standard films. The opacities shape can be round or irregular, and the number of opacities in each lung zones can be described by a discrete score from 0/-, 0/0, 0/1, 1/0, 1/1, 1/2, 2/1, 2/2, 2/3, 3/2, 3/3, and 3/+, corresponding to twelve possible categories, which is named profusion level. Based on the perfusion level of the opacities on the CXR of a patient, a stage between 0 and III is assigned to the patient to indicate the severity of the disease [2].

In China, pneumoconiosis holds ~90% occupational diseases. The number of prevalence cases of pneumoconiosis is 638,234, up to the end of 2008, while there are over 610,000 suspicious cases, reported by Ministry of Health of the Peoples Republic of China [3]. However, the diagnosis based on x-ray film is time consuming, subjective to staging, high experience depended.

In this paper, we present a computer aided detection (CAD) tool to detect the pneumoconiosis in digital Posterior-Anterior (PA) chest radiographs by image analysis automatically, which follows the process of lung segmentation, feature extraction in lung fields and sub-zones, classification and chest based report out. The CAD tool will help the radiologists in screening and staging applications as the

second view, which will improve the work flow in mass screening, reduce the work load in x-ray images reading, and keep a consistent accuracy in pneumoconiosis diagnosis.

## **2 Lung Segmentation and Sub-division**

Segmenting the lungs from digital chest x-ray image is to get the lung fields for the following automatic analysis. Also, the automatic segmentation of lung fields from chest radiographs can be a useful tool for the morphology analysis based detection of abnormalities.

### **2.1 Active Shape Model Segmentation**

Lung field segmentation on chest radiographs has attracted a number of researchers. Rule based schemes have been investigated by Ginneken [4]. Ginneken and Romeny proposed a new hybrid segmentation scheme that combined the strengths of a rule-based approach and of a pixel classification approach. Vittitoe developed a pixel classifier for the identification of lung regions using Markov random field modeling. Ginneken et al. put forth a lung field algorithm based on ASM method with optimal features. Iglesias investigated lung segmentation based on the detection of oriented edges and active contours models. As the efficiency and accuracy are required in the segmentation process, ASM algorithm is employed to delineate the lung fields in this paper.

The ASM algorithm could be mainly divided into two steps, training and segmentation.

In the training step, a procedure is to be followed to build up the model, i.e. labeling lung contour landmarks on training images; aligning manual landmark shapes by scaling, rotating and translating; calculating the statistics of a set of aligned shapes; constructing the shape model.

The second step is to delineate the lung contour for an input image. The procedure includes initializing the process with the average shape obtained from training phase; searching the optimal point along the direction perpendicular to the contour for each landmark in the initial contour; finding the best fitted lung field shape and update the parameters and the shape model; repetition until the process converges.

Figure 1 shows the result of lung field segmentation with ASM algorithm.

### **2.2 Subdividing the Lung Fields into Six Zones**

According to the domain knowledge of pneumoconiosis diagnosis, the texture features are extracted from multiple regions to better capture the features of the diffuse abnormalities. On the basis of ILO guidance and radiologists' diagnostic measurement, six non-overlapping regions are subdivided in the lung fields. Each region will be analyzed with a separate classifier that is built by the features extracted solely from this region. The goal is able to capture the feature of abnormal variations

in that particular part; in principle, this will reduce the influence of superimposed anatomical structures, which tend to make the abnormalities difficult to distinguish.

Given a lung field mask image, each lung field is subdivided into three zones - the upper, middle and lower zone. Right lung field subdivision is done by calculating the apex and diaphragm position and computing horizontal lines that divide the field into three parts of equal height. However, as the large variability of left lung shapes leads to difficulty in obtaining diaphragm position, the left lung field subdivision differs from the right lung field. Left lung field is subdivided by calculating the apex and bottom position and computing horizontal lines that divide the field into three parts of equal height, see Figure 2.

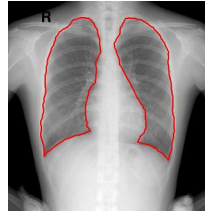


Fig. 1. The result of lung field segmentation using ASM algorithm.

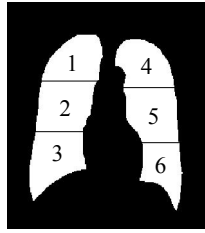


Fig. 2. The lung fields are subdivided into six regions.

### 3 Feature Extraction and Selection

Once a region of interest (ROI) – each of 6 zones – is segmented, the next task is to characterize it in terms of a set of features. Here, we are following two procedures to achieve the goal. First, we extract three types of features from each ROI, totally 247 features. Second, we select a set of features for the following classification by correlation analysis and clustering method.

#### 3.1 Feature Extraction

##### 1. Image intensity based feature

We extract a set of 6 features based on the histogram of intensity values – mean, standard deviation, skewness, kurtosis, energy and entropy. Apart from calculating

these on the original ROI, we also extract these features after applying a difference filter on the image. The purpose of a Gaussian filter is to perform a local enhancement operation. The filter is applied at various scales and orientations so that the right level of enhancement is achieved.

#### 2. Co-occurrence matrix based feature

We also extract a set of 5 features based on the gray level co-occurrence matrix computed for the ROI, namely energy, entropy, local homogeneity, correlation and inertia. The co-occurrence matrix allows us to capture the level of similarity and dissimilarity among adjacent pixels in an ROI. Thus, an ROI with opacity will contain adjacent pixels with similarly high intensities, whereas a normal ROI will not contain such adjacent pixels. Computing these features for various orientations captures this information for various types of adjacency.

#### 3. Frequency domain based feature

We also extract a set of 5 features based on a frequency domain transformation of the ROI. These features are the mean, RMS, M1, horizontal filter and vertical filter. In total, there are 247 features extracted as described above. Given the number of data points used for model building, this number is too large and needs to be pruned. We shall discuss the methods used for feature down-selection in the following subsection.

### 3.2 Feature Selection

We select a subset of features for classification by analyzing similarities (such as correlation) and eliminating redundancy. The objective is to retain the features that achieve maximum compression of the data with minimum information loss. However, the real measure of the selection method is simply whether the set of features provides a good enough description of the ROIs to allow for a good model to be built to classify ROI into twelve categories of profusion level, which is described in section 1. Therefore, we build classifiers with feature subsets of various sizes, and choose the smallest feature set where the classifier performance is acceptable.

Consider a data set with  $m$  rows (each representing an ROI) and  $n$  columns (each representing a feature that describes the ROI). We can now create an  $n \times n$  matrix of values, where the value in cell  $(i, j)$  represents the similarity (absolute value of correlation) between feature  $i$  and feature  $j$  in the dataset.

If we consider each feature as a “point” in some space and the appropriate entry in this matrix as the similarity between pairs of points, we can cluster these points to arrive at groups of similar features. We perform clustering of features using a method known as hierarchical clustering [5]. First, features are grouped into a dendrogram, which is a tree where elements close to each other are clubbed together into branches. Depending upon the criterion provided by the user (number of clusters, distance between clusters etc), one can split the tree into branches and thus arrive at clusters. We can use this method to arrive at groups of similar features from the given dataset. Once these groups are given, we choose one feature from each group as a representative.

## 4 Classification

From machine learning point of view, pneumoconiosis detection is a binary classification problem, where classification is to determine a subject has or not has pneumoconiosis. This is the purpose of screening for workers under high risk of dust. On the other hand, pneumoconiosis staging, to determine the stage (or severity) of a patient's pneumoconiosis, is a multi-class classification problem. This can be regarded as the activity for quantitative diagnosis. This stage information coming out of pneumoconiosis staging is crucial because the determined stage is used as both legal evidence for workman's accident compensation and the means for selecting appropriate treatments.

### 4.1 Classification for Pneumoconiosis Screening

Normally, the task of building a classifier based on the available data can be viewed in terms of two major sub-tasks: finding the optimal level of complexity for the classifier, and building the classifier at the specified level of complexity.

For screening application, we build up 8 classifiers for left/right lungs, and 6 zones by sub-division. The methodology used for building the models is described in [6], and is outlined here:

1. Search through the space of possible parameter values that describe classifier complexity.
2. For each possible set of parameter values, build and evaluate the performance of the classifier. We perform this evaluation using the k-fold cross validation method, whereby a dataset is divided into k parts,  $k-1$  of which are used for training and the  $k^{th}$  is used for validation. This is done in k different ways, and the average performance across these validation cycles is reported as the performance of the classifier at that level of complexity.
3. Pick the parameter values where the best performance is achieved and build the final classifier using those values. The final classifier performance is evaluated using a method known as leave-one-out validation, which works the same way as k-fold validation but with k equal to the number of rows in the dataset. The leave-one-out performance is considered to be a good measure of the generalization ability of a classifier [7].

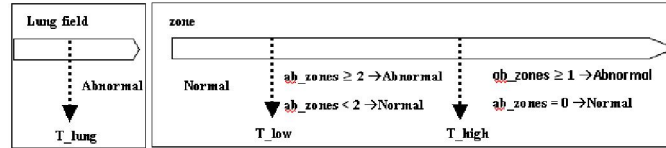


Fig. 3. The strategy of chest classification from two levels.

Once models for lung and zone level ROIs have been built, we evolve the chest based final report by leveraging the domain knowledge of pneumoconiosis screening, for example, the classification probability of one lung ROI is larger than a threshold,



the chest will be labeled as abnormal, but for zone ROIs, the final conclusion is generated from how many abnormal zones and how much the abnormal zone's probability. The strategy is shown in figure 3. The thresholds are optimized by training in database.

For all the individual classifiers, we use Support Vector Machines (SVM), a machine technique pioneered by Vladimir Vapnik [7, 8] and now extensively used in a number of applications. An SVM attempts to construct a hyperplane in a high dimensional space that has the largest distance (maximal margin) to the examples in the training set close to the hyperplane. Complexity is scaled by transforming the original features using an appropriate kernel function.

A key design choice to be made in building SVMs is the selection of the kernel to compute the similarity between two examples. We use the Radial Basis Function (RBF) kernel in our case, and control the complexity by varying the gamma( $\gamma$ ) parameter. The other parameter we control is the penalty for the error term  $C$ . These two parameters determine the complexity of the classifier being built. In order to choose the right level of complexity as defined by these two parameters, we measured the performance on unseen examples of an SVM model built at a particular complexity level using the k-fold validation technique.

## 4.2 Classification for Pneumoconiosis Staging

The pneumoconiosis detection problem has been actively studied. However, work on pneumoconiosis staging is very sparse, in spite of its importance.

Since the pneumoconiosis stages (0, I, II, III) are discrete categories ordered by disease severity, to be more exact, the pneumoconiosis staging problem is an ordinal classification problem. Ordinal classification is a special machine-learning task that falls between classification and regression [9]. Unlike nominal classification, where the target classes are un-ordered and classification performance is measured by classification accuracy, for ordinal classification both accuracy and the distance between the actual and predicted classes are relevant in evaluating classification performance. Hence, an important design strategy in developing ordinal classifiers is to leverage the ordering information to minimize the distance between actual and predicted classes.

In literature there are several approaches towards addressing ordinal classification problems. The simplest approach would be to solve ordinal problems by treating them as un-ordered nominal classification problems. This "passive" approach ignores the ordering information that could be used to improve the classification performance. More sophisticated approaches have been proposed in recent years, for example, [10-13].

In this paper, we propose using SVM for pneumoconiosis staging. In our model, the inputs are the selected features and outputs are the pneumoconiosis stages. To address the ordinal classification problem, we explore several design strategies, including categorical, ordinal, and regression.

## 5 Experiments

In this section, we introduce the experiments by applying the algorithms described above on a database of digital chest X-Ray images comprising normal and pneumoconiosis chest x-ray images.

Normal chest images and pneumoconiosis chest images were collected for building up the database. The normal chest images are categorized by gender and age, while the pneumoconiosis chest images are categorized by stage, from 0+ to III.

A total of 857 images were collected, of which 632 were of normal subjects and 225 were of patients afflicted with pneumoconiosis. And we randomly select training set from the database, including 252 normal, 175 abnormal; test set includes 380 normal and 50 abnormal images.

The algorithm was tested on the database mentioned above. Ground truth for segmentation validation was available on 400 images from this database. The performance of the algorithm was evaluated on the basis of DICE similarity coefficient (DSC) between the segmented lung mask and the ground-truth mask. DICE coefficient is computed as twice the area overlap (logical AND) of the two masks divided by the sum of the areas of the masks (logical OR). On the 400 cases tested, 89% of the images have  $> 85\%$  DSC, and about 96% cases have  $DSC > 80\%$ .

Of the 247 features extracted at each resolution, a smaller subset was selected for classification. At the lung level, the correlation threshold-based selection method was employed with a threshold of 80% to select 38 features. At the lung zone level, the hierarchical clustering method was used to select between 12 ~ 14 features for each zone. 2 models were built at the lung level (one for each lung) and 6 at the zone level (one for each zone). The outputs of these models were combined to get the final result for the chest according to the strategy in figure 3.

We used the LIBSVM package to build and test the SVM classifiers that were used in the ensemble. This package uses a variant of the Sequential Minimal Optimization Algorithm [21, 22] to build the SVM. An RBF kernel was used in the SVM, and the  $\gamma$  and  $C$  parameters which determine classifier complexity were both varied in the range  $[2^{-15}; 2^{15}]$  in order to arrive at the model with optimal complexity and good generalization. Table 1 gives a summary of the results in the training and test sets.

**Table 1.** Experimental results in training and test sets.

	Sensitivity	Specificity	Accuracy
Training set	99.5%	92.7%	95.5%
Test set	98.1%	79.6%	81.8%

## 6 Conclusions and Discussion

Pneumoconiosis detection, especially for coal workers, has attracted many researchers in the past decades [14-16]. The computer scheme we present here not only can detect lung interstitial diseases by lung texture analysis, but also provides the possibility of quantitative diagnosis. Currently it is expected to provide value in

occupational disease detection, including mass screening and quantitative staging applications. However, it should be possible to extend it to interstitial diseases with diffuse character, such as lung tuberculosis.

## References

1. Dehong Li, Physicians textbook of occupational diseases pneumoconiosis. Publishing House of Peoples Daily, Beijing (2004)
2. ILO, Guidelines for the use of the ilo international classification of radiographs of pneumoconiosis - revised edition, International Labor Office, (2000)
3. Xinhua News Agency, Report on Ministry of Health of the Peoples Republic of China, Beijing (2009)
4. Ginneken BV, Romeny BM, Viergever MA: Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on Medical Imaging* 20(12): 1228-1241 (2001)
5. Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley-Interscience (2000)
6. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, A practical guide to support vector classification, <http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>
7. Vladimir Vapnik, *Statistical Learning Theory*, John Wiley (1998)
8. Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press (2000)
9. S. Kramer, G. Widmer, B. Pfahringer and M. DeGroeve, "Prediction of ordinal classes using regression trees", *Fundamenta Informaticae* (2001)
10. McCullagh, P., "Regression model for ordinal data", *Journal of the Royal Statistical Society Series B* (42) 109-142 (1980)
11. J.R. Potharst and J.C. Bioch, "Decision trees for ordinal classification", *Intelligent Data Analysis*, Vol. 4(2), pp. 97-112 (2000)
12. Chu & Keerthi. "New approaches to support vector ordinal regression", *Proc. 22nd International Conference on Machine Learning*, Bonn, Germany (2005)
13. Mathieson, M. J. "Ordinal models for neural networks", in *Neural Networks in Financial Engineering*. Eds. A. P. Refenes, Y. Abu-Mostafa and J. Moody. World Scientific, Singapore (1995)
14. R. Kruger, W. Thompson, and A. Turner, "Computer diagnosis of pneumoconiosis," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-4, 1974.
15. A. Turner, R. Kruger, and W. Thompson, "Automated computer screening of chest radiographs for pneumoconiosis," *Investigation Radiology*, vol. 11, pp. 258--266, 1976.
16. Hiroshi Kondo and Takaharu Kouda, "Computer-aided diagnosis for pneumoconiosis using neural network," in *14th IEEE Symposium on Computer-Based Medical Systems (CMBS'01)*, 2001.