FINAL REPORT - GRANT 1 R01 OH03055-01

AUTOMATED STAGING OF COAL WORKERS' PNEUMOCONIOSIS UTILIZING DIGITAL IMAGE ANALYSIS OF CHEST X-RAYS

by

Kendall Preston, Jr., (Principal Investigator)

Erik Kurt Witt (Graduate Student)

Electrical and Computer Engineering Department
and
Biomedical Engineering Program

CARNEGIE MELLON UNIVERSITY
5000 Forbes Avenue
Pittsburgh, PA 15213

Phone: (412) 268-2521
Fax: (412) 268-1173

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO.	2.	PB97-147268

4. Title and Subtitle Automated Staging of Coal Workers' Pneumoconiosis Utilizing Digital Image Analysis of Chest X-Rays	5. Report Date 1996/02/26
	6.
7. Author(s) Preston, K., Jr., and E. K. Witt	8. Performing Organization Rept. No.
9. Performing Organization Name and Address Electrical and Computer Engineering Department and Biomedical Engineering Program, Carnegie Mellon University, Pittsburgh, Pennsylvania	10. Project/Task/Work Unit No.
	11. Contract (C) or Grant(G) No. (C) (G) R01-OH-03055
12. Sponsoring Organization Name and Address	13. Type of Report & Period Covered
	14.

15. Supplementary Notes

16. Abstract (Limit: 200 words) This final report covers research designed to determine the feasibility of automated staging of coal workers' pneumoconiosis using digital image analysis of chest X-rays. A system which can potentially aid the radiologist in the classification of CWP from chest X-rays is described. The system applies statistical pattern recognition to digitized CWP chest X-ray images based on the generation of features from the matrix of numbers that represents the digitized films, and the application of a variety of algorithms that may be used to distinguish profusion and opacity size. Both traditional (2-D Fourier transform and the spatial gray level dependence matrix) and new (minimum tracking algorithm and the Gauss-Markov random field parameters) feature sets were used; 386 features were included. Comparison of an expert radiologist's classifications with those of other radiologists yielded correlation coefficients of 0.60 and 0.82 for profusion and opacity size, respectively. When the expert's classifications were compared to the computer classifications, the correlation coefficients were 0.56 for profusion and 0.55 for opacity size. The authors conclude that ultimately the classifier and visualization tools described can be used in a preliminary chest X-ray screening system which would identify regions possibly exhibiting pathology, and indicate these regions to the radiologist.

17. Document Analysis a. Descriptors

b. Identifiers/Open-Ended Terms NIOSH-Publication, NIOSH-Grant, Grant-Number-R01-OH-03055, End-Date-11-30-1995, Pulmonary-system-disorders, Diagnostic-techniques, Coal-workers-pneumoconiosis, Respiratory-system-disorders, Lung-disease Occupational-respiratory-disease

c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages 15
	22. Security Class (This Page)	22. Price

INTRODUCTION

This is the final report on National Institute for Occupational Safety and Health grant number 1 R01 OH03055-01 to the Biomedical Engineering Program, Carnegie Mellon University. This report covers research designed to determine the feasibility of automated staging of coal workers' pneumoconiosis using digital image analysis of chest X-rays.

There are between 150 and 200 thousand individuals in the United States who are exposed to coal dust, primarily from working in underground coal mines. These workers are likely to develop CWP (coal workers' pneumoconiosis) over the course of their working lifetimes. This disease is debilitating, and, in order to prevent the disease from causing significant health problems, every coal worker is given the opportunity to receive a free chest X-ray examination at five year intervals. Currently, about 8,000 screening chest films are taken per year in the United States, but this number has been as high as 24,000 in the 1970's. [1] These films are collected at the NIOSH (National Institute of Occupational Safety and Health) repository in Morgantown, West Virginia. To date, approximately 375,000 such films have been accumulated. [2]

Radiologically, CWP presents primarily as rounded opacities in standard posterior anterior (PA) chest X-rays. In order to classify CWP, the International Labour Organization [3] (ILO) provides a set of standard films and guidelines. In the ILO system, the main parameter for staging the disease is *profusion*, which is an indication of the *number of opacities per unit area.* A 4-point profusion scale is defined, which ranges from 0 (normal) to 3 (maximum disease stage). the major categories on this scale can be divided into 3 subcategories each, for a total of 12 minor profusion categories. A parameter for opacity *size* is also represented in the ILO standard films. The 3-point size scale is defined as: p (0-1.5mm), q (1.5-3mm), and r (3-10mm). This scale can be subdivided into 7 minor size categories. It is very difficult to define opacity boundaries, however, so individual opacities cannot be measured or counted. The determination of disease stage is based completely on a *qualitative* visual comparison with the ILO standard set of films.

Unfortunately, the visual analysis or *reading* of CWP films is extraordinarily difficult with the result that inconsistent results are obtained. Although a great deal of effort has been put into the creation of standards for the staging process, and training of radiologists, there continues to be considerable variability among radiologists in staging the disease.

In order to illustrate this problem, two matrices are given comparing, on a given set of films, the agreement between readers of the same film. [4] Table 1.1 gives the probability of the first reader profusion category given the second reader profusion category. This data clearly indicates that there is much disagreement in staging profusion between readers. For example,
$P(R = 1 \mid R2 = 1) = 0.39$. Only 39% of the time does the first reader agree with the second on films exhibiting the lowest pathological category of profusion. Even more important are false negatives, assuming the second reading is the truth:
$P(R1 - 0 \mid R2 = 1) = 0.52$. 52% of category 1 profusion films as staged by the second reader are considered normal by the first.

Table 1.1

Probability of Reader 1 Major Profusion Category
Given Reader 2 Major Profusion Category

		R 2		
	Category	0	1	2 - 3
R	0	0.91	0.52	0.09
1	1	0.08	0.39	0.24
	2 - 3	0.01	0.09	0.66

R1= A-readers (screening radiologists)
+ B-readers (NIOSH certified)
R2= B-readers only

When agreement by ± 1 minor profusion category (each profusion category is divided into 3 minor categories for a total of 12) is considered, the results in Table 1.2 are obtained. Agreement beyond normal plus one minor profusion category (0/1) is poor.

Table 1.2

Reader Agreement for ± 1 Minor Profusion Category

Minor Category	% Agreement
0/0	97
0/1	98
1/0	27
1/1	30
1/2	16
2/1 +	24

In certain cases, such as 1/1 profusion, there is more disagreement than agreement among radiologists. This is extraordinarily significant in that the coal miner is directly affected by the radiological stage of his chest film. The screening of all coal miners for the development and progression of CWP is vital, in that with a positive chest X-ray, a miner has the option to be relocated to an area of lower air dust content before the risk of developing potentially disabling respiratory conditions becomes significant. Furthermore, in claims for disability compensation, as well as epidemiological studies employing large numbers of films, it is essential that each film is read without bias and with consistency over time. The value of an effective screening program cannot be disputed in light of the disabling pulmonary conditions that can arise from an occupation so vital to

our industries and economy. The high cost of modern medical care is, or course, another major consideration.

In order to attack this problem, there have been many attempts in the USA, Europe, and Japan to apply computer pattern recognition to digitized images of CWP films. The studies which developed multi-profusion stage classifiers are listed in Table 1.3.

Table 1.3

Prior Work (Multi-Profusion Stage Studies)

Authors	Year	Institute	Training Set		Test Set	
			# Films	Accuracy	# Films	Accuracy
Desaga	1987	U. Gießen.Ger.	250		200	82%
Hall	1975	Yale	38	85%		
Kruger	1974	USC/LA	leave 1 out train/test		95	50-66%
Kobatake	1987	Tokyo, Japan	11	42-61%		
Jagoe, Patton	1975	Middlesex, Engl.	36	80%		
Jagoe	1979	"	36		96	55%
Chen, Toriwaki	1990	U. Nagoya, Japan	7 - 11	50% -88%		

see Table 1.4 for more details

Considerable prior work has been performed on the topic of automatically staging CWP. Many of these studies report excellent classification results in controlled experiments. However, *none of this research has lead to a practical system which would benefit the people with or at risk for pneumoconiosis.* One system, for example, was actually developed for NIOSH, and transferred to Morgantown for testing. None of the results obtained by the originating organization could be reproduced—the equipment could not be made to perform in the manner in which it was claimed to be capable.

Therefore, the research reported was undertaken in order to advance studies on the application of digital image processing to the analysis of CWP films. This work is a cooperative effort with the National Institute of Occupational Safety and Health (NIOSH, Morgantown, WV) which compiles coal workers' diagnostic and epidemiological statistics for the United States, performs epidemiological studies, and administers the Coal Workers' X-Ray Surveillance Program through which the film repository is maintained. First, an end-to-end system was developed, including a presentation of *all steps* involving input data collection, feature generation, classifier training and testing, and data visualization. Second, non-ideal films were used in training. These films were selected at random from the NIOSH repository, which contains films obtained *on location* for screening coal workers. The data set includes films from the late 1960's and 1970's with variable exposure and little or no calibration of the X-ray machine (variable keV films). Third, optimal feature sets for the discrimination of normal films from pathological films were determined using both linear and quadratic classifiers. Fourth, the quadratic classifiers incorporated three parameters in classification: profusion, size, and spatial location. Fifth, visualization tools were developed which can be incorporated into a radiologist's display system to provide additional diagnostic information beyond a simple stage provided by the classifier.

PROCEDURE

The two major steps taken in applying statistical pattern recognition to digitized CWP chest X-ray images are (1) the generation of features from the matrix of numbers that represents the digitized film, and (2) the application of a variety of algorithms in finding those features or combinations of features that may be used in distinguishing profusion and opacity size. In the research undertaken here, traditional features were selected that had been used in prior research and additional features were added or devised that, it was believed, had certain unique properties that might make them valuable. In contrast to prior research, the number of features (386) was far greater than every previously employed. The feature classes that were utilized are:

Table 1.4

Traditional and New Feature Sets Employed in Classifier Training

Traditional Feature Sets	Total
2-D Fourier Transform	**40**
ring energy sums (various types of normalization)	24
first moment of energy spectrum in bands	16
Spatial Gray Level Dependence Matrix (SGLDM)	**80**
statistics	40
spatial gray level dependence vector statistics	40

New Feature Sets	Total
Minimum Tracking Algorithm (Gray level sizing algorithm)	**48**
Gauss-Markov Random Field Parameters (2-D linear prediction)	**208**

About 120 tiles (32x32 pixel subimages) were selected from the inter-rib spaces from each film used in this project. For each tile, from every x-ray, a large feature vector is computed containing those features listed in Table 1.4

Two types of classifiers were utilized: linear classifiers and quadratic classifiers. The linear classifier requires the calculation of the quantity given by

$$F = (\mu_1 - \mu_2)^T (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2) \qquad (1.1)$$

where μ_1 and μ_2 are the class feature mean vectors, and Σ_1 and Σ_2 are the class covariance matrices for all tiles from the films in training classes 1 and 2, respectively. This quantity (F) is computed for each feature subset considered in classifier training. Once the optimal feature set is known (greatest F), a feature weight vector ($\mathbf{w}$) is computed using

$$\mathbf{w} = (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2) \qquad (1.2)$$

In order to use this classifier in the assignment of a disease stage to a tile selected from the input x-ray, Equation 1.3 is used,

$$p(\mathbf{x}) = \mathbf{w}^T\mathbf{x} - w_0 \qquad (1.3)$$

where $\mathbf{x}$ represents an N-dimensional feature vector, and w_0 a scalar offset. This equation (with $p = 0$) defines a hyperplane in an N-dimensional space. The number computed (p) indicates by its polarity whether the point being classified lies to one side or the other side of this hyperplane. If the point lies on one side, it is classified as class 1, on the other, class 2. By testing the data for a specific film using several such equations, the identity of the film can be assigned to a particular degree of profusion and opacity size. In the research reported here, three classes were employed in training: normal (N); high profusion small opacity (P); and high profusion large opacity (R). These three training classes were used because they represented extremes in both profusion and size.

In this work, we determined the optimal set of features to use in classification by exhaustive search of all possible feature subsets drawn from a large pool. Traditional methods, such as stepwise linear discriminant analysis, suffer from two major disadvantages: First, the optimal feature set for classification can be easily missed by this training method—the optimal classification set may not include the best single feature for classification, or even members of the best classification feature pair, for example. In fact, a test of traditional stepwise linear discriminant analysis returned *none* of the features discovered by exhaustive search in the optimal 3 feature classifier. Second, the output of this procedure is a single classifier. It is often the case that more than one set of classifiers stands out from the set of all possible classifiers. Moreover, studying the best *set* of classifiers may provide insight into the classification problem at hand.

This leads to a further problem, namely, the computational effort necessary for the exhaustive search of large feature spaces. Table 1.2 shows the computer

time required for searching feature spaces having one through five dimensions using an 80486-based machine, and a dedicated Cray C90. Note that with quadratic classifiers, discussed later, these numbers increase by nearly *4 orders of magnitude.* It is evident that with the method of exhaustive search, using more than three dimensions is too computationally intensive. This is why our work was limited to three dimensions.

Table 1.5: Computer Time Required to Perform Linear Classifier Training from a 368 Feature Pool (FLOP's only)

Classifier Dimension	Feature Combinations	CPU Time (seconds)	
		80486/DX2 50	Cray C90
1	3.68E+2	3.2E-2	1.9E-4
2	6.75E+4	5.9E+0	3.5E-2
3	8.23E+6	7.1E+2	4.3E+0
4	7.52E+8	6.5E+4	3.9E+2
5	5.47E+10	4.7E+6	2.8E+4

The difficulty in employing separating planes, i.e., linear separation, for CWP data is indicated in Figure 1.1.

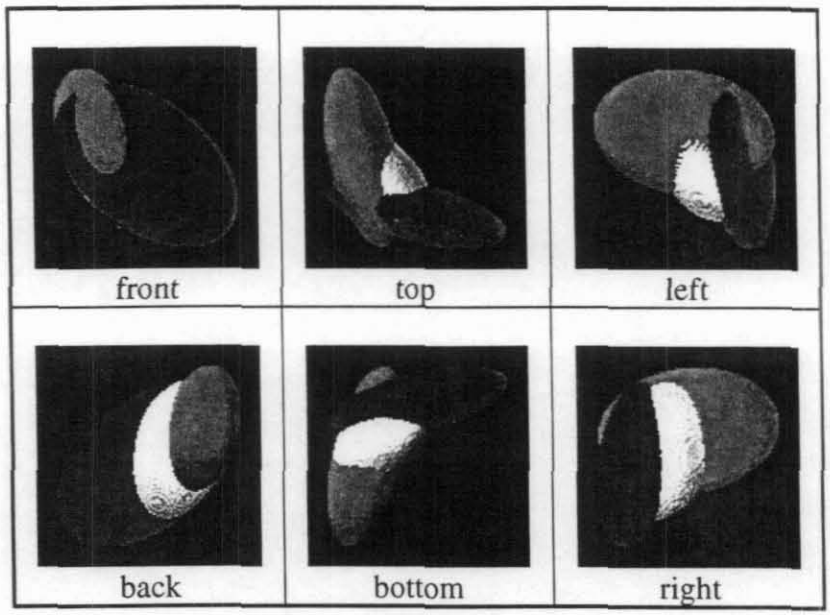

Figure 1.1: Example probability ellipsoids for the three training classes (*N* [Green], *P* [Yellow], and *R* [Red])

This figure shows the nearest fitting ellipsoids to the three compartments selected in a 3-dimensional feature space: normal (N), high profusion p-type (P), and high profusion r-type (R) training film sets. Because of the curvilinear nature of the data in each compartment, a better-suited classification paradigm utilizes the quadratic classifier, with which we continued our studies. Using the quadratic classifier, assignment is performed using Equation 1.4,

$$C_n \ni \|\mathbf{x}-\mu_n\|^2_{\Sigma_n^{-1}} = \min_m\left(\|\mathbf{x}-\mu_m\|^2_{\Sigma_m^{-1}}\right), \qquad (1.4)$$

where

$$\|\mathbf{x}-\mu_n\|^2_{\Sigma_n^{-1}} = (\mathbf{x}-\mu_n)^T \Sigma_n^{-1}(\mathbf{x}-\mu_n) \qquad (1.5)$$

defines the Mahalanobis distance from class n, where μ_n and Σ_n are the class feature mean vectors and covariance matrices, respectively and $\mathbf{x}$ is the tile feature vector. Equation 1.4 states that the decision rule is to assign to the tile the class to which it has the smallest Mahalanobis distance. In this case the polarity of the number calculated is always positive and represents a normalized distance from a particular data point to the centroid of an ellipsoid whose shape, location, and orientation is determined from the covariance matrices of the reference data for each of the three extrema.

RESULTS

Using this approach, *data from all three classes were uniquely and correctly classified.* In order to classify the intermediate disease stages, the method of *classification signatures* was developed. A classification signature is the percent of a film's tiles classified into the three disease extremes (the three compartments in feature space: N, P and R). The percentage of tiles classified as high profusion (P + R) was shown to be correlated with profusion as assigned by the expert radiologist. Likewise, the difference in percent of tiles assigned to each opacity size category (R - P) was shown to be correlated to opacity size as assigned by the expert radiologist.

Table 1.6 summarizes these correlations, and compares them with similarly obtained correlation coefficients using other radiologists' stages for the same films. The correlation coefficients for profusion are nearly identical for the computer and radiologists. The opacity size correlation for computer prediction is less than that for human agreement; but a large part of this finding can be attributed to the limitation in size categories used by the other radiologists. It is important to note that profusion is directly related to the actual amount of coal

dust in the lung. This type of relationship has not been demonstrated for opacity size. Therefore, profusion is the most important factor in classifier output. In light of these facts and findings, the quadratic classifier appears to perform as well as a typical reader.

Table 1.6 Correlation Coefficients Comparing Radiologists' Performance with the Computer

	Expert Radiologist vs.	
	other radiologists	computer
Profusion	0.60	0.56
Opacity Size	0.82*	0.55

**other radiologists used 3-point rather than 7-point size stage*

Now that classification results from the optimal 3-feature quadratic classifier have been presented, it is useful to examine *properties* of the selected features for insight into why they are the optimal set. Because the x-ray quality of the training film set has a wide range, the classifier should be robust to background gray level, and to a certain extent, gray level modulation.

The first 2 (of 3) features used in the classifier are Fourier energy ring sums normalized by AC energy. These features are computed from the 2 most-inner rings centered around the DC point of the Fourier transform image. The first detects structures larger than 1.8mm, and the second, structures from 1.0mm to 1.8mm. Both of these features are therefore expected to be sensitive to larger opacities. These features are from the group of Fourier features most resistant to film quality changes for the following reasons: The background gray level (DC) is subtracted before performing the Fourier transform. And, they are normalized by all AC energy (the sum of energy in all the rings, not including DC energy). Therefore, they measure the relative low-frequency energy (not absolute energy), and are resistant to changes in modulation.

The third feature in the classifier is a texture feature. The matrix is generated for a spacing of 2 pixels in the vertical direction. The specific statistic is the gray-level *correlation* of all pixel pairs in the defined spatial orientation. A high value for this feature indicates that the change in image gray-level is relatively small over short distances. This feature, like the others in the classification set, is immune to changes in background gray-level (additive), and gray-level modulation (multiplicative). The fact that a vertically oriented feature was chosen does not imply directional sensitivity, since no direction-independent feature was available for the classifier training program to choose as an alternate.

In this report, a system has been described which can potentially aid the radiologist in the classification of coal workers' pneumoconiosis (CWP) from chest x-rays. The foundations of a solid classifier have been presented from which a clinical system to aid in diagnosis can ultimately be developed. Visualization tools are a major component of this system. These tools are used to indicate to the radiologist potential regions of pathology which require further investigation. Ultimately, the classifier and visualization tools could be used in a preliminary chest x-ray screening system which would identify regions possibly exhibiting pathology, and indicate these regions to the radiologist superimposed on an image of the chest film.

FUTURE RESEARCH

In light of the encouraging results obtained from this work, it is worthwhile to discuss the direction of future research on this topic in order to develop a system which can be used clinically. First, in order to address the major deficiency in the research that has been conducted with Morgantown to date, all films utilized should be examined by a panel of expert radiologists rather than a single radiologist. Every member of the panel should be tested in order to obtain data on the reproducibility of their conclusions made from visual examinations. Variation in readings may explain some of the difficulty encountered by the classifier in staging intermediate films. Second, repeat digitizations should be

made; and the variability due to the digitizer should be determined. Third, we see no reason to expand the feature set used in training. All classes of features typically used in pattern recognition, which apply to this type of problem, are included in the training feature set. Fourth, the ILO staging system is a radiological description—it is not necessarily a predictor of outcome in terms of the individual's health. It may be more useful to train the classifier based on outcome. Output from this classifier could be used to predict the most effective treatment, and determine fair compensation. Fifth, better training data could improve classification accuracy. As noted previously, many of the training films are 20 or more years old. Although the modern chest x-ray is by no means ideal, the techniques and instrumentation have improved over the past two decades. Also, digital imaging systems now being developed (e.g., using phosphor plate technology) will improve image quality significantly. Sixth, processing time continues to be an issue in classifier training. Linear classifiers require a few minutes to train (parametric distribution vs. distribution), and quadratic classifiers almost a day (data vs. parametric distribution). The next logical step is a nearest neighbor approach (data vs. data), which does not assume a parametric distribution. In fact, this technique was coded for, but it was determined that it would take three quarters of a year to complete on a fast linear machine. As the performance of parallel processors improves, this processing time will become very manageable in the near future.

Finally, this system is not restricted to the automated staging of pneumoconiosis. It is a very general system which could be easily applied to other medical or general imaging problems. Areas in which texture differences provide information on diagnosis, such as mammography, are potential candidates for the algorithms described and implemented in this dissertation.

ACKNOWLEDGEMENTS

The research reported here was supported primarily by Grant 1 R01 OH03055-01 and was also supported by the Section of Informatics at the University of Pittsburgh under National Library of Medicine grant 5 T15 LM07059 that also provided office space and computing facilities during 1993-1995. The

assistance of Dr. Carl Fuhrman, Chief of Thoracic Radiology at the University of Pittsburgh Medical Center; Dr. Jack Parker, Chief, Examination Processing Branch, NIOSH; Dr. John Hankinson, Chief, Clinical Investigations Branch, NIOSH; Dr. Greg Cooper, Director, Medical Informatics Training Program, University of Pittsburgh; Dr. Richard Stern, Dr. Virginia Stonick, all of the Department of Electrical and Computer Engineering, Carnegie Mellon University, are also acknowledged.

REFERENCES

1. U.S. Department of Health and Human Services (NIOSH): Work relaed lung disease surveillance report. [Based on 1982-1987 data.] 91-113:36, Oct,1991.

2. Hankinson, J.L.: written communication, April 1992.

3. International Labour Organisation: Guidelines for the use of ILO international classification of radiographs of pneumoconioses. No. 22 (rev. 80), occupational safety and health series. Geneva . International Labour Office, 1980.

4. Althouse, R.B., Attfield, M.D., Hodous, T.K.: Inter-reader variability among readers using ILo 1971 and 1980 classification of the pneumoconioses. Transactions: VIIth International Pneumoconioises Conference, 1434-1439, Aug. 1988.