35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Fully Automated Scoring of Chest Radiographs in Cystic Fibrosis

Min-Zhao Lee, Weidong Cai, *Member, IEEE*, Yang Song, *Student Member, IEEE*, Hiran Selvadurai and David Dagan Feng, *Fellow, IEEE*

Abstract— **We present a prototype of a fully automated scoring system for chest radiographs (CXRs) in cystic fibrosis. The system was used to analyze real, clinical CXR data, to estimate the Shwachman-Kulczycki score for the image. Images were resampled and normalized to a standard size and intensity level, then segmented with a patch-based nearest-neighbor mapping algorithm. Texture features were calculated regionally and globally, using Tamura features, local binary patterns (LBP), gray-level co-occurrence matrix and Gabor filtering. Feature selection was guided by current understanding of the disease process, in particular the reorganization and thickening of airways. Combinations of these features were used as inputs for support vector machine (SVM) learning to classify each CXR, and evaluated using two-fold cross-validation for agreement with clinician scoring. The final computed score for each image was compared with the score assigned by a physician. Using this prototype system, we analyzed 139 CXRs from an Australian pediatric cystic fibrosis registry, for which texture directionality showed greatest discriminating power. Computed scores agreed with clinician scores in 75% of cases, and up to 90% of cases in discriminating severe disease from mild disease, similar to the level of human interobserver agreement for this dataset.**

I. Introduction

Cystic fibrosis is one of the most common life-threatening genetic disorders worldwide, affecting up to 1 in 2500 people born in the highest-risk populations [1]. This prevalence corresponds to a total frequency of 1 in 25 of recessive alleles which encode defective chloride ion channels in epithelial cells. The disease causes considerable morbidity and mortality, affecting multiple organs and ultimately with an average life expectancy at birth of close to 40 years despite ongoing medical care [2].

Although cystic fibrosis is characterized by the formation of cysts and fibrotic scar tissue within the pancreas, the most severe consequence of the disease is its impact on lungs, where impaired chloride transport leads to thick mucus production and inability to clear these secretions from the lungs. Bacteria proliferate in the favorable environment, establishing a sequence of chronic infection with frequent acute exacerbations [3].

The course of disease in lung follows a typical course, with progressive inflammatory thickening of the airways and destruction of alveoli. Early changes most commonly affect the upper regions of the lung before becoming apparent elsewhere. Subsequently, airways become dilated, and this together with trapping of mucus leads to airflow obstruction. Later thickening of airways and vessels tends to progress from central to peripheral regions. Destruction of lung architecture can also increase the resistance of pulmonary blood circulation, leading to pulmonary hypertension and consequently cardiac failure.

Clinicians assess the severity of disease with a combination of measures including patient symptoms, physical examination, pulmonary function tests, sputum culture and radiological imaging. These assessments are used in documenting disease progression, guiding interventions, evaluating the response to treatment, and predicting mortality. Substantial effort has been devoted to comparing the different measures of disease severity, in terms of their predictive value and correlation with other measures [4, 5, 6].

Plain chest radiographs (CXRs) are a significant component of the assessment of cystic fibrosis in children, and several different scoring systems exist to quantify the degree of abnormality seen. Different systems are better-established in each of three major regions with relatively high prevalence: Shwachman-Kulczycki scoring in Australia [7], Chrispin-Norman scoring in Europe, and Brasfield (Birmingham) and Wisconsin scoring in North America. All scoring systems refer to the visible lung changes associated with disease progression, albeit with slight differences in focus. In particular, clinicians look for signs of airflow obstruction (expanded shape of the chest cavity), bronchial and vascular thickening (linear markings), nodules and cysts, and gross regional abnormalities.

Shwachman-Kulczycki scoring classifies CXRs into five categories, from 25 to 5 in steps of 5, in order of increasing severity. Table I describes the CXR findings for each score, as initially proposed by Shwachman and Kulczycki. A numerical scoring system for CXRs with defined features of interest is a particularly attractive target for automation, but work to date in this area has been very limited. Scoring of CXRs for cystic fibrosis patients is still performed entirely by clinicians. An automated scoring system would provide clinicians with an objective measure of the CXR changes.

In this study, we examine the use of a segmentation algorithm to identify lung fields, from which we extract textural features proposed by Tamura, local binary patterns (LBP), gray-level co-occurrence matrix (GLCM) features, and Gabor filter outputs. We then combine these features to classify CXR images with a final score of 10, 15 or 20, and evaluate the overall performance of the automated system in terms of its agreement with the scores given by clinicians. Scores of 5 and 25 were grouped with 10 and 20 respectively, as too few (only 3 of 139) cases were assigned such scores.

Research supported in part by the Australian Research Council.

M. Lee is with the School of Information Technologies, University of Sydney, Australia, and Royal Prince Alfred Hospital, Sydney, Australia.

W. Cai, Y. Song and D. D. Feng are with the School of Information Technologies, University of Sydney, Australia.

H. Selvadurai is with the Sydney Medical School, University of Sydney, Australia, and Sydney Children's Hospitals Network, Sydney, Australia.

978-1-4577-0216-7/13/$26.00 ©2013 Crown

TABLE I. SHWACHMAN-KULCZYCKI X-RAY SCORING

Points	Findings
25	Clear lung fields.
20	Minimal accentuation of bronchovascular markings; early emphysema.
15	Mild emphysema with patchy atelectasis; increased bronchovascular markings.
10	Moderate emphysema; widespread areas of atelectasis with superimposed areas of infection; minimal bronchial ectasia.
5	Extensive changes with pulmonary obstructive phernomena and infection; lobar atelectasis and bronchiectasis.

II. MATERIAL AND METHODS

A. Acquisition of CXR Data

Patients with cystic fibrosis were identified from an Australian pediatric cystic fibrosis registry. Patients undergo yearly reviews, which include a scoring CXR. Prior to 2009, CXR images were taken with different protocols and stored in several different formats, rendering them too variable for our analysis. Of 279 patients, 139 (aged 2 to 16) had undergone a sufficiently recent pediatric scoring CXR study which was available for analysis. The most recent study was selected for each patient, with the corresponding score documented in its report at the time. A clinician reviewed the final 139 images for technical problems that were likely to impact the analysis; none were excluded on this basis though note was made of considerable variation in patient positioning and beam penetration. As part of this review, the same clinician scored the CXRs independently, allowing the measurement of interobserver agreement for this dataset. The clinician manually segmented the lung fields in each images at its original size, by tracing lung outlines and thereby defining lung masks comprised of every pixel within the outlines.

B. Preprocessing

The CXR images were resampled and intensity-normalized for both segmentation and feature analysis. For automated segmentation, original images (ranging from 721 × 696 to 1131 × 951 pixels) were downsampled to 256 × 256 pixels using nearest-pixel interpolation. Intensity normalization was performed by histogram stretching, mapping the 25th and 75th centiles to intensities of 0.25 and 0.75 respectively. Manually segmented lung masks were also downsampled to 256 × 256 pixels by the nearest-pixel method. For texture analysis, original images were downsampled to 512 × 512 pixels using linear interpolation.

C. Automated Segmentation

Automated segmentation of each image was performed and evaluated as a two-fold cross-validation, using a patch-based nearest-neighbor mapping algorithm (Figure 1), which used the manual segmentation of images in the other fold as reference maps.

In each 256 × 256 image, 16 × 16 pixel patches were extracted at horizontal and vertical intervals of 8 pixels. Each such target patch was then compared to nearby (within 3 intervals) patches in other images. The distance between each pair of patches was computed as the Pythagorean sum of all corresponding pixel value distances:

$$d^2(L, M) = \Sigma_{x,y} (L_{x,y} - M_{x,y})^2 \qquad (1)$$

where L and M are the patches being compared, and x and y are the coordinate offsets within each patch.

Segmentation of each target patch was then determined by pixel-wise voting between its 11 nearest neighbors, relying on the assumption that patches of similar appearance and location are likely to have similar segmentation. For each target patch, its 11 nearest neighbors (selected as a balance between voting accuracy and computational complexity) were aggregated, and pixels which were manually segmented as lung in a majority of these, would be segmented as lung in the target patch. Similarly, pixels mostly segmented as background in the reference patches would be considered background in the target patch.

D. Texture Analysis and Scoring

Lung fields were divided into 4 regions corresponding to left and right upper and lower zones, a partitioning used in the Chrispin-Norman scoring system [8]. Lungs were divided into upper and lower halves by area in pixels. Textures within each region, and across the image overall, were analyzed using Tamura features [9], LBP [10], GLCM properties [11] and Gabor filters [12].

Tamura features of coarseness, directionality and contrast were computed for each 32 × 32 pixel block within the corresponding lung masks, and aggregated for each region. These features correlate well with visually perceptible differences on CXR (Figure 2). A separate LBP histogram and GLCM was computed for each region. GLCM features exhibit some correlation with similar Tamura features,

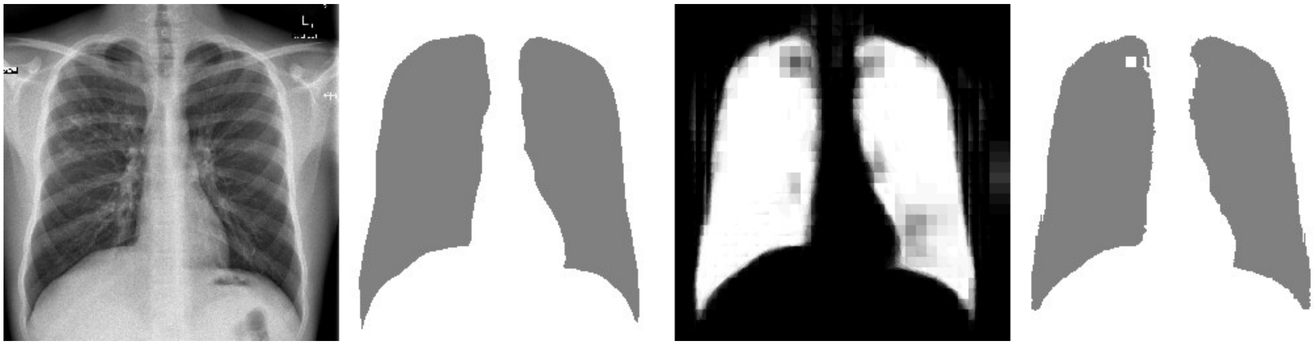

Figure 1. Example of lung segmentation, on the image with the median overlap. From left to right: downsampled and normalized chest radiograph image; manually segmented mask; aggregated votes; final machine segmentation.

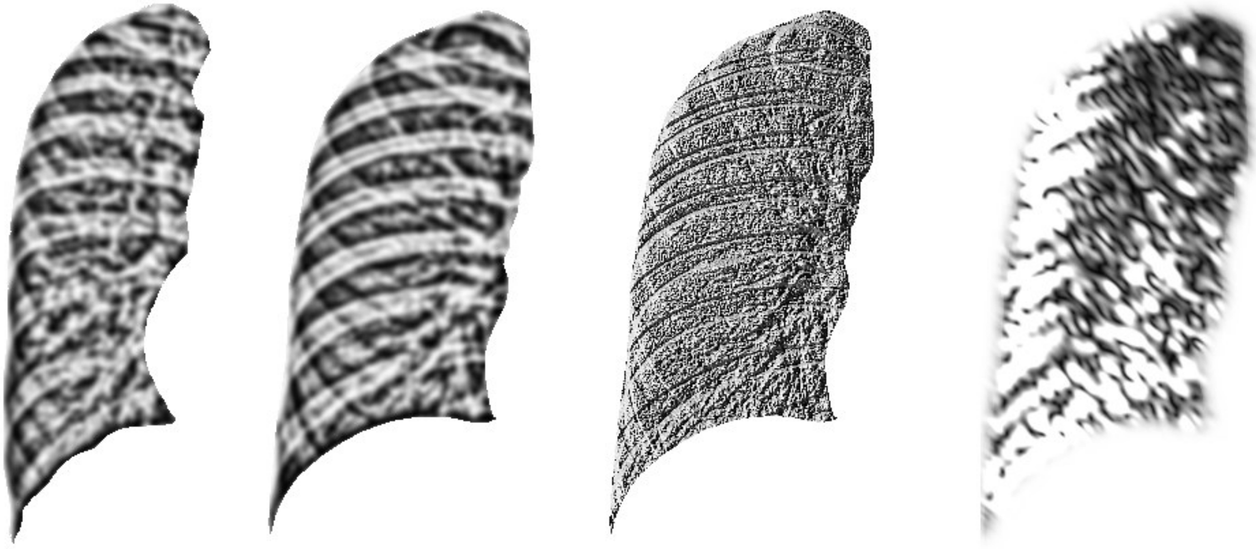

Figure 2. Texture analysis. From left to right: comparison of lung texture on chest radiographs between severe disease (score 10, far left) showing disorganised vascular and bronchial changes; and mild disease (score 20) with normal vascular appearance; local binary pattern image for lung field with mild disease; and Gabor filtered image (far right) for lung field with mild disease showing strong response in the lower half corresponding to vessels running diagonally, and confounding signal from rib tips on the left edge.

although they have fallen somewhat out of favor in comparison to LBP. LBP provides some advantages in being relatively intensity and contrast invariant, making it robust in the face of changes in beam penetration. Gabor filtering was performed in 4 directions at 3 scales over each entire normalized 512 × 512 pixel image, with the filter response averaged for each region. As with Tamura features, these filters would be expected to identify coarseness and directionality, including the specific direction of markings in each lung region.

The output vectors of each of these methods were concatenated across regions, then used individually and in combination as inputs for a SVM learning model [13]. The performance of the classification using SVM was evaluated using two-fold cross-validation, across the entire dataset, and again using only the subset of images (total 84) with scores of 10-or-below (severe) and 20-or-above (mild). Performance was compared with the interobserver agreement for clinicians on the same data.

III. Results and Discussion

A. Automated Segmentation

The performance of automated segmentation was validated using overlap,

$$\Omega = TP / (TP + FP + FN), \qquad (2)$$

where true positive (TP) is the correctly segmented lung area, false positive (FP) is the actual background area classified as lung, and false negative (FN) is the actual lung area classified as background.

The median overlap between automated segmentation and manual segmentation was 0.939, with 25th and 75th centiles at 0.921 and 0.950, respectively. This performance is comparable to other previously proposed methods of automated segmentation and close to the interobserver overlap between human observers [14]. The authors of this previous study, which dealt with images of good technical quality and few gross abnormalities, discussed the limitations of manual segmentation fairly extensively in their report.

In contrast, patients with moderate or severe CF often manifest large variations in CXR appearance, on a variety of scales (including cystic changes on a smaller scale, consolidation or collapse on a larger scale, and overall morphological changes to the thoracic wall). Furthermore, although variation between adults is already considerable, the variation across the developmental period from 2 to 16 years can be expected to be far greater.

Despite these challenges, patch-based mapping maintained a high level of performance, demonstrating its robustness for this application. A similar approach could conceivably be applied to segmentation or identification of other features on any radiograph of a standard body region, including the heart, or air-fluid levels on abdominal x-ray, or even tissues on volumetric imaging.

The patch-based mapping method encountered certain difficulties also reported previously [14], most commonly the inclusion of a stomach bubble as part of the left lung field. This remains a problem for any similar method that does not explicitly specify shape information.

Computation time, using a C++ implementation on 2.26 GHz Intel Core i5 and 4 GB RAM, was 30 seconds per image for segmentation with a reference set of 70 images.

B. Texture Analysis and Scoring

The vast majority of CXRs (136 of 139) had clinician assigned scores between 10 and 20. This is the typical distribution of scores in the CF population. Scores outside this range were grouped with the closest score within the range, giving the three classifications 10 (severe), 15 (moderate) and 20 (mild). Clinician interobserver agreement was 0.70 across all three classifications, and 0.95 when only severe and mild groups were considered.

Total image directionality provided the best machine classification performance, at 0.75 for three classifications and 0.90 for severe-mild discrimination. Introducing coarseness, and contrast, alone or in combination, did not appreciably affect this result.

Directionality compared very favorably to all other textural features (Table II), notwithstanding the increased computational cost incurred in calculating the GLCM or Gabor filtering the image. All features demonstrated markedly higher performance in severe-mild discrimination, compared to classification of the entire dataset.

This performance was comparable to the agreement between human observers for this dataset, as well as the median in historical studies [15]. Nevertheless, it appears that scoring of CXRs in disease is a more difficult task than either lung segmentation or identifying gross patterns of disease.

Part of this difficulty can be ascribed to the subtlety of changes along the spectrum of disease, which even human observers are frequently unable to distinguish if presented with small blocks of CXR images in isolation. A patient with a moderate overall score could have localized severe changes interspersed with lung tissue of relatively normal appearance. For this reason, the average textural features seem to be a better measure than the values for individual blocks. Indeed, it has been shown that even 2nd order texture characteristics may be insufficient to describe the visible differences [16].

Further complicating the task is the fact that disease severity follows a normal, rather than multimodal, distribution. Existing scoring systems are more qualitative than quantitative in nature, and in this way the classification of images into score-labeled groups is a somewhat artificial distinction. The difficulty in discriminating CXRs in the middle range of severity is borne out by the markedly improved performance when the task is transformed from a three-way classification problem to a two-way classification problem ("severe" or "mild" only).

TABLE II. AGREEMENT WITH CLINICIAN-ASSIGNED SCORE FOR EACH FEATURE INPUT SET CONSIDERED

Feature Inputs	Agreement, for classification with scores 10, 15, 20	Agreement, for classification as severe / mild only
total image directionality	0.75	0.90
Tamura	0.75	0.86
clinician observer	0.70	0.95
LBP + Tamura	0.66	0.85
(all computed features)	0.66	0.83
LBP	0.60	0.86
LBP+GLCM	0.60	0.83
LBP+Gabor	0.56	0.85
Gabor+GLCM	0.53	0.64
Gabor	0.51	0.76
GLCM	0.46	0.64

CONCLUSION

We have introduced a new fully automated scoring system for chest radiographs (CXRs) in cystic fibrosis. The system uses automated segmentation and analysis of texture features, to achieve overall agreement of up to 75% with clinician scores. In distinguishing severe disease from mild disease, the system achieves performance of above 85%, and up to 90% using a measure of image directionality.

This prototype system offers several opportunities for further development, including the use of multiscale patch-based mapping segmentation, clustering of patches, the use of new and higher-order texture features, contextual features, and alternative machine learning algorithms. Future work could potentially correlate image features with other clinical measures of disease severity such as symptoms and signs, functional capacity, and pulmonary function tests.

REFERENCES

[1] F. Ratjen, G. Döring, "Cystic fibrosis," in *Lancet*, vol. 361, Feb. 2003, pp. 681-9.

[2] J. R. Yankaskas, B. C. Marshall, B. Sufian, R. H. Simon, D. Rodman, "Cystic fibrosis adult care: consensus conference report," in *Chest*, vol. 125 1 Suppl, Jan. 2004, pp. 1S-39S.

[3] S.M. Rowe, S. Miller, E. J. Sorscher, "Cystic fibrosis," in *N. Engl. J. Med.*, vol. 352, May 2005, pp. 1992-2001.

[4] M. Corey, V. Farewell, "Determinants of mortality from cystic fibrosis in Canada, 1970-1989," in *Am. J. Epidemiol.*, vol. 143, May 1996, pp. 1007-17.

[5] S. Terheggen-Lagro, N. Truijens, N. van Poppel, V. Gulmans, J. van der Laag, C. van der Ent, "Correlation of six different cystic fibrosis chest radiograph scoring systems with clinical parameters," in *Pediatr. Pulmonol.*, vol. 35, June 2003, pp. 441-5.

[6] R. H. Cleveland, D. Zurakowski, D. M. Slattery, A. A. Colin, "Chest radiographs for outcome assessment in cystic fibrosis," in *Proc. Am. Thorac. Soc.*, vol. 4, Aug. 2007, pp. 302-5.

[7] H. Shwachman, L. L. Kulczycki, "Long-term study of one hundred five patients with cystic fibrosis," in *AMA J. Dis. Child.*, vol. 96, July 1958, pp. 6-15.

[8] A. R. Chrispin, A. P. Norman, "The systematic evaluation of the chest radiograph in cystic fibrosis," in *Pediatr. Radiol.*, vol 2, 1974, pp 101-5.

[9] H. Tamura, S. Mori, T. Yamawaki, "Textural Features Corresponding to Visual Perception," in *IEEE Trans. on Systems, Man, and Cybernetics*, SMC-8, June 1978, pp. 460-472.

[10] T. Ojala, M. Pietikäinen, T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, July 2002, pp. 971–987.

[11] R. M. Haralick, K. Shanmugam, I. Dinstein, "Textural Features for Image Classification," in *IEEE Trans. on Systems, Man, and Cybernetics*, SMC-3, Nov. 1973, pp. 610–621.

[12] I. Fogel, D. Sagi, "Gabor filters as texture discriminator," in *Biol. Cybernetics*, vol. 61, 1989, pp. 103–113.

[13] C. Cortes, V. Vapnik, "Support-vector networks," in *Machine Learning*, vol. 20, Sep. 1995, pp. 273-297.

[14] B. van Ginneken, M. B. Stegmann, M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database", in *Med. Image Anal.*, vol. 10, Feb. 2006, pp. 19-40.

[15] N. Lewiston, R. Moss, R. Hindi, S. Rubinstein, M. Sullivan, "Interobserver variance in clinical scoring for cystic fibrosis", in *Chest*, vol. 91, June 1987, pp. 878-82.

[16] B. Julesz, E. N. Gilbert, J. D. Victor, "Visual discrimination of textures with identical third-order statistics", in *Biol. Cybernetics*, vol. 31, Sep. 1978, pp. 137-40.