Automated estimation of progression of interstitial lung disease in CT images

Yulia Arzhaeva[a)]
CSIRO Mathematical and Information Sciences, New South Wales 1670, Australia
and Image Sciences Institute, University Medical Center Utrecht, 3508 GA, The Netherlands

Mathias Prokop
Department of Radiology, University Medical Center Utrecht, 3508 GA, The Netherlands

Keelin Murphy and Eva M. van Rikxoort
Image Sciences Institute, University Medical Center Utrecht, 3508 GA, The Netherlands

Pim A. de Jong and Hester A. Gietema
Department of Radiology, Meander Medical Center, Amersfoort, 3800 BM, The Netherlands

Max A. Viergever and Bram van Ginneken
Image Sciences Institute, University Medical Center Utrecht, 3508 GA, The Netherlands

(Received 18 May 2009; revised 14 September 2009; accepted for publication 27 October 2009; published 4 December 2009)

Purpose: A system is presented for automated estimation of progression of interstitial lung disease in serial thoracic CT scans.
Methods: The system compares corresponding 2D axial sections from baseline and follow-up scans and concludes whether this pair of sections represents regression, progression, or unchanged disease status. The correspondence between serial CT scans is achieved by intrapatient volumetric image registration. The system classification function is trained with two different feature sets. Features in the first set represent the intensity distribution of a difference image between the baseline and follow-up CT sections. Features in the second set represent dissimilarities computed between the baseline and follow-up images filtered with a bank of general purpose texture filters.
Results: In an experiment on 74 scan pairs, the system classification accuracies were 76.1% and 79.5% for the two feature sets, respectively, while the accuracies of two observer radiologist were 78.5% and 82%, respectively. The agreements of the system with the reference standard, measured by weighted kappa statistics, were 0.611 and 0.683 for the two feature sets, respectively.
Conclusions: The system employing the second feature set showed good agreement with the reference standard, and its accuracy approached that of two radiologists. [DOI: 10.1118/1.3264662]

Key words: computer-aided diagnosis, interstitial lung disease, computed tomography, serial CT evaluation, texture analysis, dissimilarity

I. INTRODUCTION

Interstitial lung disease (ILD) is a chronic inflammation of the lung parenchyma, encompassing over 150 specific disorders causing significant morbidity and mortality.[1,2] In recent years, computed tomography (CT) has received a central role in the diagnostics of ILD.[3,4] ILD manifests itself in CT images as a variety of abnormal patterns in the lung parenchyma. Clinical estimation of disease progression is based on monitoring changes in those patterns, along with the results of physiologic tests. A change in the visual extent of disease over time is an important marker of response to therapy and a predictor of mortality. In this work, we propose an automated system for assessment of interval changes in ILD based on quantitative measurements in serial CT images.

Clinical literature conventionally agrees that an increase in the overall extent of parenchymal abnormalities is associated with disease progression and a decrease with disease regression. Disease extent is estimated visually by a radiologist. Although highly specialized chest radiologists show moderate agreement, most CT scans with ILD are interpreted by general radiologists who might provide less reproducible and accurate results. With the introduction of multidetector CT that allows one to obtain near volumetric CT scans, the amount of image data a radiologist has to go through in order to compare two scans has increased considerably. This makes the observational estimation of disease progression a time-consuming task. A recent study[5] showed that the times required to assess ILD changes in a scan pair were on average of 123 and 79 s for a pair of nonaligned and aligned CT scans, respectively. The same study reported a fairly low intra- and interobserver agreement (Fleiss' κ=0.54 and 0.58 for nonaligned and aligned pairs, respectively). Therefore, the automation of ILD progression assessment is a clinically valuable computer-aided diagnosis (CAD) application that can offer reproducible and fast estimates of disease changes.

One possible approach to automation would be the automated estimation of the overall extent of parenchymal abnormalities computed as the accumulated extent of different abnormal patterns. Recently, considerable advances have been made by the medical computer vision community in the field of texture classification in lung CT. In high resolution CT, regions of interest from two-dimensional (2D) axial sections were automatically classified into several texture categories representative of ILD (see Refs. 6–8, for a review). Automatic classification of 3D volumes of interest (VOIs) from multidetector CT scans showed a very good reproducibility of the reference standard set by expert radiologists.[9,10] There is a gap, however, between classification of independent regions and estimation of disease extent in the whole scan or an axial section.

Not many attempts to bridge this gap have been made so far. Reference 6 addressed the question whether a given CT section contained a certain abnormal pattern. In Ref. 10, four complete lung volumes were manually annotated, and these annotations were compared with the output of an automated classification system that classified every small VOI in the lungs into four texture categories. The last study showed statistically that the computer system agreed with the experts as well as the experts agreed between themselves in labeling these four subjects. To our knowledge, no work has been published that directly uses the results of automated texture classification in the estimation of change in overall disease extent.

It should be noted that no reliable reference standard yet exists for training a texture classification system. Present systems are trained on a single expert's annotations or annotations obtained by the consensus of a panel of experts. Making such annotations is a laborious and time-consuming task and leads to high intra-and interobserver variability.[6,8]

Reproducible quantitative CT measures have been investigated in studies directly related to the estimation of ILD progression in serial CT scans. In Refs. 11 and 12 simple statistical features were computed to characterize the distribution of intensities of the lung volume: mean lung attenuation, skewness and kurtosis in Ref. 12 and variance, contrast and entropy in Ref. 11. Best *et al.*[12] showed that all three features changed significantly in patients with deteriorated idiopathic pulmonary fibrosis (a clinical syndrome often associated with ILD). Sumikawa *et al.*[11] demonstrated significant differences in measurements in 13 cases of nonspecific interstitial pneumonia (a common subtype of ILD) before and after treatment.

Our study is the first attempt to automatically assess ILD progression in a patient using quantitative measurements from a pair of CT scans. Disease progression is estimated separately in the lower, middle, and upper parts of the lungs. To this end, a pair of corresponding axial sections is analyzed in each part of the baseline and follow-up scans. Alignment of two scans is an important preprocessing step that enables the retrieval of matching sections. Our system, applied to a pair of CT sections, yields an opinion whether the second image in the pair corresponds to a higher, lower, or equivalent extent of disease compared to the first image.

Two sets of quantitative features are investigated for use with an automated analysis. The first set includes statistical features which describe the intensity distribution of the difference image computed between corresponding baseline and follow-up CT sections. For the second set, we derive new dissimilarity features from local texture features that were previously shown to be able to characterize different abnormal patterns associated with ILD.[6,13] The dissimilarities between individual texture features are used to directly estimate the difference between two images, thereby skipping classification of the lung parenchyma into different abnormal categories. In this way, we avoid the laborious and unreliable step of obtaining manual annotations for training a texture classification system.

This paper is organized as follows. A data set used for training and validation of the system is described in Sec. II. Section III gives the system overview and details each part. Section IV describes the experimental setup and observer study. The results are presented in Sec. V and discussed in Sec. VI.

II. MATERIALS

II.A. Data set

Seventy five pairs of baseline and follow-up thoracic CT scans of patients with histologically proven ILD were collected from daily clinical practices of the University Medical Center Utrecht (21 pairs) and St. Antonius Hospital Nieuwegein (54 pairs), The Netherlands, between 2003 and 2007. Types of ILD included sarcoidosis, idiopathic interstitial pneumonias, and various immune and autoimmune disorders. All patients that underwent more than one CT examination in the given period of time and had a confirmed ILD diagnosis were included in the study except for those whose scans exhibited severe motion artifacts. The time span between the baseline and follow-up scans varied from 1 month to 2 years. The data set comprised 40 male and 35 female patients, with a mean age of 53 years (range: 25–77 years) at the time of a baseline scan.

Images were obtained at full inspiration on a multidetector row scanner (Brilliance-16P, Mx8000 IDT 16, Brilliance-40, or Brilliance-64, Philips Medical Systems, the Netherlands), with standard or low-dose parameters for high-resolution volumetric CT scanning. Collimation varied between 0.625 mm (40-and 64-slice scanners) and 0.75 mm (16-slice). Slices of 0.9 mm thickness (40- and 64-slice) or 1 mm thickness (16-slice) were reconstructed every 0.7 mm at the University Medical Center Utrecht. Slice thickness and spacing were 0.8 mm (16-slice) in the St. Antonius Hospital Nieuwegein. Exposure settings ranged between 15 and 180 mAs, with 120 or 140 kVp. All images had a per-slice resolution of 512 × 512, with pixel spacing in the X and Y directions varying from 0.3 to 0.8 mm.

II.B. Reference standard

Three axial sections used in the assessment of ILD progression were manually extracted from the baseline scan at the approximate distance of 2 cm above the carina (the upper part of the lungs), 2 cm below the carina (the midpart), and at 1 cm above the diaphragm (the lower part). The sequential numbers of extracted sections in the whole scan were noted. Then, three sections with the same sequential numbers were taken from the follow-up scan. Due to the previously executed image registration, the sections taken from the follow-up scan were at the same level of anatomy as the sections extracted from the baseline scan.

Corresponding pairs of CT sections were annotated by an expert chest radiologist with more than 15 years of experience. In a side-by-side comparison, the expert classified a change in disease extent in the follow-up sections. There were eight classification categories—massive decrease (disease extent reduction >50%), moderate decrease (10%–50% reduction), minor decrease (2%–10% reduction), stable (any change in the disease extent ⩽2%), minor increase (disease extent increase of 2%–10%), moderate increase (10%–50% increase), massive increase (increase >50%), or the expert could reject a pair altogether if the quality of one or both images was deemed insufficient. During the annotation process the expert had access to full CT scans, before and after registration, as well as to previous pertinent radiological reports. The expert was aware which of the two scans was baseline—its sections were always projected on the left side of the computer display.

The pairs were divided between the categories as follows: massive decrease (27 pairs), moderate decrease (17), minor decrease (11), stable (105), minor increase (24), moderate increase (20), and massive increase (1 pair). Twenty pairs of CT sections were discarded by the expert because of motion artifacts, registration misalignments, or a combination of both. Motion artifacts deform the appearance of parenchyma and obscure the difference between normal and abnormal tissues which prevents both human observers and a computerized analysis from correct interpretation. The misalignment of a patient's baseline and follow-up scans makes it impossible to extract corresponding slices in the two scans, which is a critical step in our analysis. Often, motion artifacts or misalignments affected only a part of the lungs and only sections from the affected parts were discarded. Among the 20 discarded pairs, however, 3 pairs were from the same patient, which reduced the final number of participants in the study to 74. In total, there were 205 pairs of CT sections included in the study.

The "stable" category strongly prevails over any other category in the data set, which is clinically realistic but unsuitable for training a CAD system. Such an inequality in class sizes can make classification biased toward a class that is better represented in the training set. In order to make the classification task more feasible we grouped together massive, moderate, and minor categories and defined three classes, "regression," stable, and "progression." Furthermore, we randomly selected pairs from each category to swap the baseline and follow-up images and to change the pair label to its opposite (e.g., moderate decrease to moderate increase), until we obtained the uniform distribution of massive, moderate, and minor categories in regression and progression classes. As a result, 28 cases with massive changes, 37 cases with moderate changes, and 35 cases with minor changes were equally divided between regression and progression classes: 14-14, 19-18, and 18-17 for massive, moderate, and minor changes, respectively. The final distribution between the classes was as follows: regression—included massive, moderate, and minor decrease (51 pairs); stable—as previously defined (105 pairs); progression—included massive, moderate, and minor increase (49 pairs).

III. METHODS

III.A. System overview

The proposed automated analysis generally follows the typical design of a CAD system. At first, images are preprocessed and useful discriminatory features are computed from them. This is followed by an automated analysis of patterns described by computed features. The goal of the system is to assign a pattern to the right class. The CAD system operates in two phases. In the training phase, the system, equipped with a classification function, or a classifier, learns the parameters of the classifier from a set of patterns which true classes are known. In the testing phase, the trained classifier is applied to new, previously unseen data. The system outputs either a class label or a probability for a pattern to belong to a certain class.

The preprocessing step of our system included intrapatient registration of thoracic CT scans and subsequent 3D segmentation of the lung fields. Then, three corresponding 2D sections from the upper, mid, and lower thirds of the lungs were extracted from the baseline and follow-up scans. We computed two different sets of features from pairs of 2D CT sections. Each feature set characterized a textural change between baseline and follow-up images. The first set of features statistically described the intensity distribution of a difference image obtained by subtracting the baseline image from the follow-up image. Features in the second set are dissimilarities computed between the baseline and follow-up images filtered with a bank of general purpose texture filters.

Two classification strategies were employed in the CAD system. In the first strategy, a classifier used the intensity distribution features to differentiate between three possible categories of change: regression, stable, and progression. In the second strategy, a two-stage classification was employed. In the first stage, image pairs were classified into "changed" and stable categories using the dissimilarity-based features. Pairs, that were labeled as changed, were further classified into regression or progression categories using the intensity distribution features. The CAD system, using either strategy, was evaluated by means of accuracy and weighted kappa statistics, and its performance was compared to that of two human observers.

III.B. Registration

Prior to the extraction of corresponding axial sections, the baseline and follow-up 3D scans were aligned using a two-step registration procedure. The scan to be deformed was selected randomly in each pair. First, the images were roughly aligned using an affine transformation. This was followed by an elastic deformation that allows for nonrigid lung tissue alignment. The elastic deformation was modeled by a B-spline grid.[14] During registration, a similarity between two images is maximized. For this purpose, mutual information was used as the cost function[15] in both steps. An iterative stochastic gradient descent optimizer[16] was applied. To avoid local minima, a multiresolution approach was adopted. The software package ELASTIX, version 3.90 was used. (elastix can be downloaded from http://www.isi.uu.l/elastix.)

Both registration steps involved a multiresolution strategy using a Gaussian image pyramid. For the initial affine transformation, four resolutions were used, and five resolutions were used for the nonrigid deformation. A maximum of 512 optimization iterations were performed in each resolution level during the affine transformation. For the nonrigid deformation, the optimizer performed at most 512 iterations in the first four resolution levels and 300 iterations in the last resolution level. The B-spline grid spacing used in final resolution level was eight voxels. Registration was performed on images downsampled by a factor of 2 in order to reduce the computation time. The acquired transformation was then applied to the full resolution scan. The computation time required to register one image pair was 10 min on average on a standard high-end PC, on a single core. After registration, the two scans had the same dimensionality, with comparable anatomy at the same level of sectioning.

III.C. Lung segmentation

The segmentation of the lung fields in 3D CT scans was initially performed by an implementation of a conventional lung segmentation algorithm.[17,18] This fully automatic algorithm exploits a rule-based approach to find the trachea, from which the bronchi and lungs are grown. After the trachea and main stem bronchi have been removed from the grown lung region, the left and right lungs are labeled using another set of rules, whereupon 3D hole filling and morphological closing are applied to each lung field separately.

Although generally reliable and fast, this algorithm has limitations. Relying on the assumption of contrast in attenuation between the lung parenchyma and the surrounding tissue, it tends to undersegment the lung fields in scans containing high density pathology (as often occurs with ILD). Occasionally, the algorithm was not even able to find the trachea to start segmentation with. This happens when the appearance of the trachea does not meet the assumptions made by the algorithm.

Therefore a multiatlas segmentation-by-registration approach (MAS) was applied to scans where the conventional approach failed. An atlas is an image with a known segmentation, that is, registered to a test image. Several studies have shown MAS to be a powerful segmentation tool.[19,20] MAS assumes that N atlases are registered to an image at hand (target image), resulting in N transformations from atlases to the target image. Then, N segmentation masks of the target image are obtained by applying corresponding transformations to the atlas segmentation masks. The final segmentation of the target is obtained by the pixelwise majority voting, i.e., a voxel is assigned to the final segmentation mask if it belongs to at least $N/2$ transformed atlas segmentation masks.

In the end, the lung segmentation masks of baseline and follow-up scans were merged by the "union" operation, and the resulting mask was used in the subsequent computation of features.

The lung segmentation was also performed on downsampled scans in our study. The conventional lung segmentation method, that took on average 55 s per scan, failed in 28 CT scans out of 150. Five scans where the lungs were successfully segmented by the conventional method served as atlases for MAS. Registration parameters for MAS were the same as for the intrapatient registration. Computation time for MAS was approximately 50 min per scan. It should be noted that such time-consuming operations as intrapatient registration and lung segmentation can be performed beforehand, e.g., simultaneously with an image acquisition, thus, not compromising the computation time of the CAD system.

III.D. Features from difference image

Progression of ILD is associated with the extension of abnormal patterns in a scan. Abnormal patterns typically increase the opacity of lung parenchyma and therefore lead to higher density values than normal parenchyma. This motivated us to extract discriminatory features to describe interval changes in ILD from the difference image between the aligned follow-up and baseline scans. With no change in the disease state between the two images, this difference image should not exhibit much intensity variation in the lung fields. Ideally, one would expect the intensity histogram of the lung fields of the difference image to have a large symmetrical peak around zero and a small standard deviation. If the disease state has changed over time, we would expect a biased intensity histogram—toward positive numbers in the case of disease progression and toward negative numbers in the case of regression. This is illustrated in Fig. 1. In this figure, the difference images corresponding to regression [Fig. 1(a)] or progression [Fig. 1(c)] show more darker or brighter regions, respectively, than the difference image of a stable case [Fig. 1(b)].

The difference image of the lung fields can be described statistically in a number of ways. We used four statistical features: three quartiles and the mean. The three quartiles were the 25th percentile, the median, and the 75th percentile. Prior to computing the features, the difference image was filtered with the Gaussian filter at a scale of 2, in order to decrease spurious registration discrepancies between the two images. The areas outside the lung fields were masked out and were ignored in the computation of the features.

III.E. Dissimilarity-based features

Features that only consider overall parenchyma density changes simplify what happens during development of ILD. Not only the amount but also the type of abnormal patterns can change. For example, a typical sign of deterioration is the substitution of ground glass opacities by fibrotic tissue.[21]

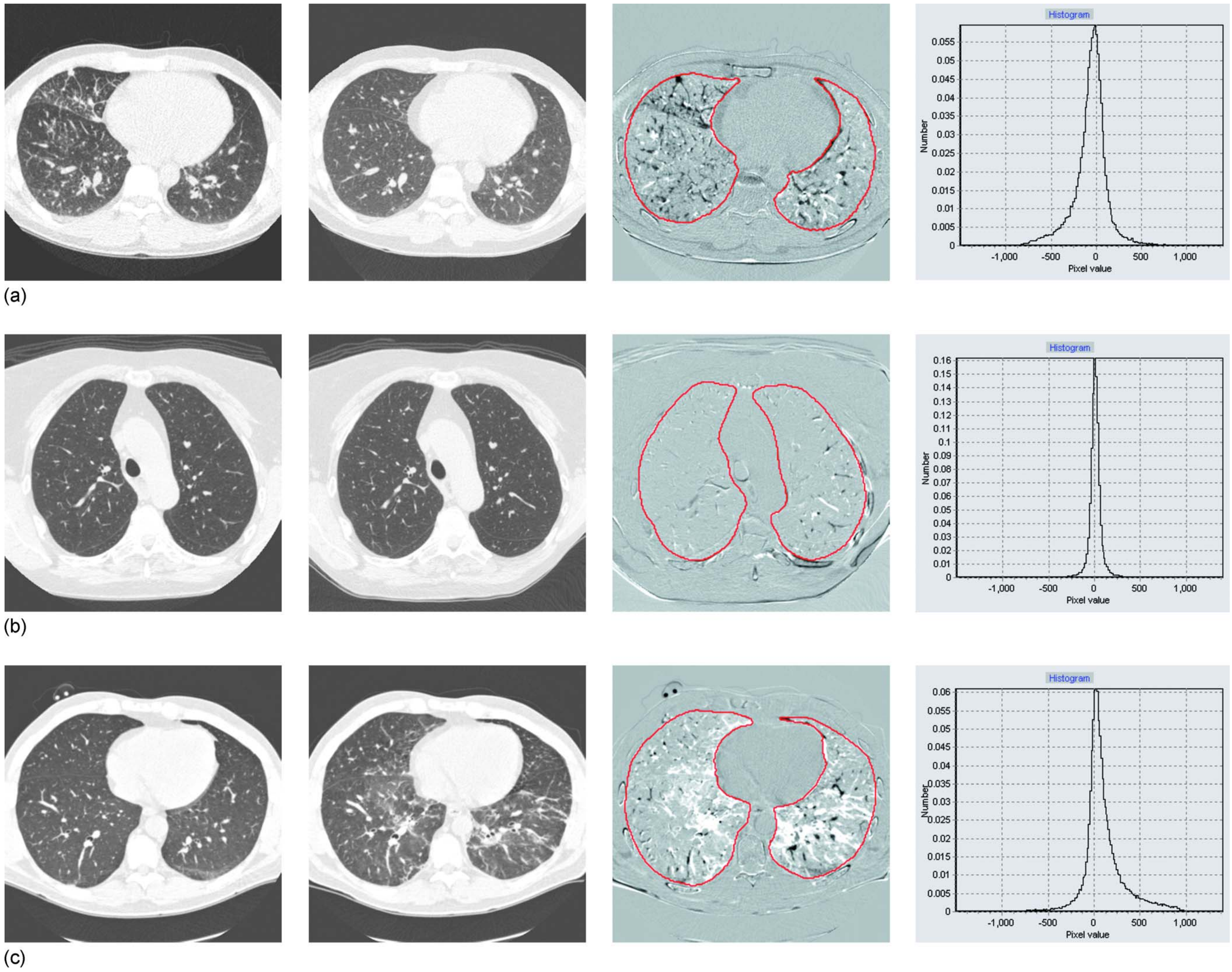

FIG. 1. Examples of difference images and their intensity histograms. In each row, from left to right: a pair of two corresponding registered CT sections of the same patient taken at different moments of time; difference image obtained by subtracting the first CT section from the second one; normalized histogram of the difference image. The histogram was computed from the lung fields only, which are outlined in the difference image. Prior to computing the histogram, the difference image was smoothed with a Gaussian ($\sigma=2$). Here the original difference images are shown.

Such changes in patterns do not always result in an increase in parenchymal opacity. On the other hand, there may always be density variation in the difference image with causes unrelated to the progression of ILD, for example, different levels of inspiration, imperfect registration, different signal-to-noise ratios in the baseline and follow-up scans, etc. As a result, image pairs with subtle changes in abnormal patterns are likely to be confused with stable image pairs with common disparities if features extracted from them only characterize the distribution of intensities in the difference image. Figure 2 shows three cases with a change in the extent of ILD where the histograms of their difference images are very similar to the typical histogram of a stable case. To better illustrate parenchymal changes related to ILD, corresponding texture patches taken from the baseline and follow-up images in Figs. 2(a) and 2(b) are enlarged and presented in Fig. 3.

Following these considerations, we propose to extract and compare the texture contents of the images in order to describe interval changes in ILD. We start with a number of texture features computed locally throughout the lung fields. Then, the distribution of each feature is described by the histogram, separately in the baseline and follow-up images. Finally, a measure of dissimilarity between the two histograms of each texture feature is computed. In this way, a pair of images is characterized by a vector of dissimilarities in texture. In the next subsection we describe the utilized texture features, and, subsequently, we detail how the dissimilarities were computed.

III.E.1. Local texture features

We used a set of general purpose texture features that have been previously applied to the classification of abnormal texture patterns in high resolution thoracic CT.[6,13] These features were four central statistical moments in eight filtered versions of the original image, calculated on three scales. The eight filters were the Gaussian, the Laplacian, the first

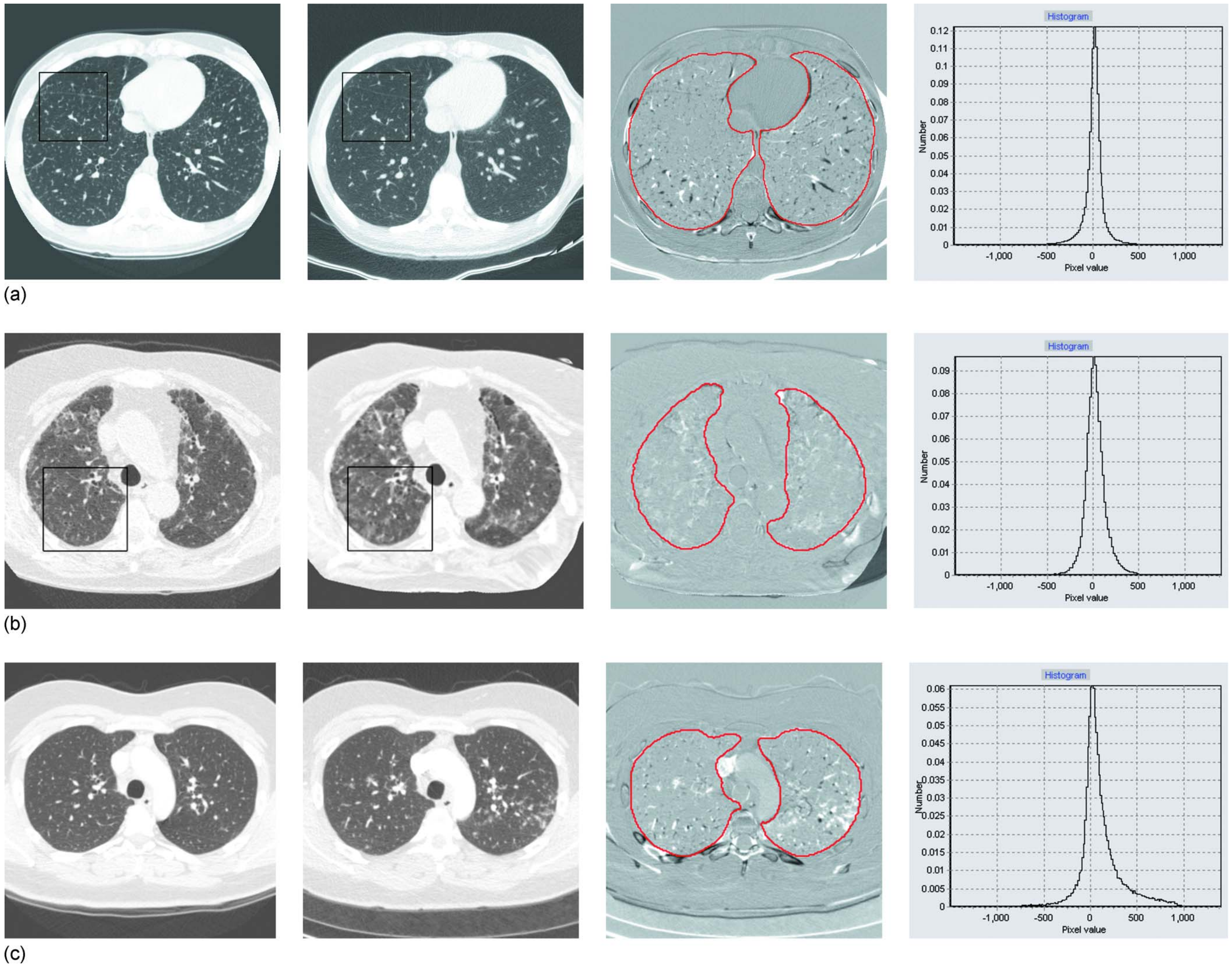

FIG. 2. Image pairs with changes in ILD extent whose difference image intensity histograms are similar to that of a typical stable case. The corresponding square patches of texture marked in (a) and (b) are enlarged and presented for comparison in Figs. 3(a) and 3(b), respectively.

order derivative of the Gaussian in three orientations between 0 and π, and the second order derivative of the Gaussian in the same orientations. The scales σ were 0.5, 1, and 2 pixels. Prior to filtering the image, pixel values in the lung fields were mirrored outside the lungs symmetrically with respect to the lung borders. This prevented a major distortion in the filter output near the lung borders which is normally caused by a large difference in appearance inside and outside the lungs. The first four moments, i.e., the mean, standard deviation, skewness, and kurtosis, were calculated from multiple regions of interest (ROIs), placed in the lung fields, in order to capture the local texture information. We used an 8×8 pixel spacing to define the centers of circular ROIs, each of which had a radius of 16 pixels. On average, there were 870 ROIs per image. In total, 96 features (8 filters $\times 3$ scales$\times 4$ moments) were computed for each ROI.

III.E.2. Computation of dissimilarities

Each texture feature distribution was represented by a 64-bin normalized histogram, which resulted in 13–14 entries per bin, on average. The bin partitioning was determined on a set of training images by computing the range of values for each feature and splitting this into 64 equal intervals.

A number of comparison measures between two distributions have been proposed by the image retrieval community (see, for example, Ref. 22, for a review). They are conventionally termed dissimilarity measures, and we adhere to this term in this paper. Let us denote the histograms of feature f, computed from images A and B, as $h_f^A=\{h_f^A(i)\}$ and $h_f^B=\{h_f^B(i)\}$, respectively, i being a bin index. A dissimilarity measure between h_f^A and h_f^B is denoted $d(h_f^A,h_f^B)$. Then, a comparison measure between A and B is obtained as a vector of dissimilarities $D(x,y)=\{d(h_f^A,h_f^B)\}$, where f is a running index through all available features.

In this study, the dissimilarity vector D of the same dimensionality as the number of local texture features was computed between the baseline and follow-up images, and this was used as a feature vector entering the second classification strategy, detailed below.

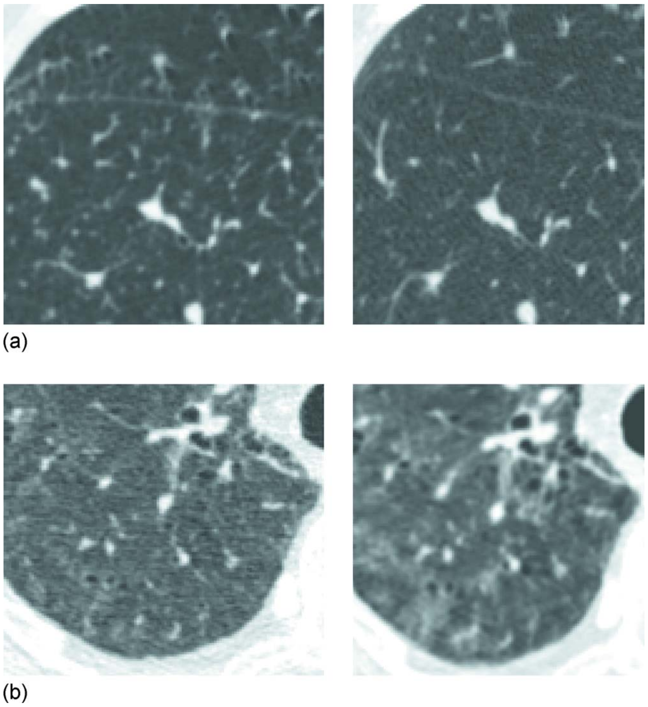

FIG. 3. Corresponding patches of texture in the baseline and follow-up CT sections. In (a), textural changes associated with ILD regression are shown (left to right) and in (b) are textural changes associated with ILD progression. These are examples of subtle changes.

III.F. Classification

III.F.1. Strategy I

Our first classification approach exploited the intensity distribution features computed from difference images. A classifier was trained to distinguish between three classes of temporal change: regression, stable, and progression. For convenience, we call a pair of the baseline and follow-up images a "training sample" when it is used by the automated system to train the classifier and its correct class is known to the system. A "test sample" is a pair that is new to the system, and the system attempts to define its class using a classification rule it has learnt.

III.F.2. Strategy II

As noted in Sec. III E 2, the dissimilarity-based features are intended to reflect how "far" from each other two images are, without specifying the "direction" of change. These features only enable the distinction between changed and stable pairs of images. To be able to perform a three-class classification with them, we developed the following strategy. First, the dissimilarity-based features were used with a two-class classifier that distinguishes between changed and stable pairs of images. Next, we applied another two-class classifier to those pairs that were labeled as changed by the first classifier. The second classifier was trained to discriminate between regression and progression cases. The same features computed from difference images as used in strategy I were used with the second classifier of strategy II.

We assume that the introduction of the dissimilarity-based features will improve the separation of changed and stable pairs, based on the considerations introduced in Sec. III E. The intensity distribution features insufficient on their own for three-class classification, play a role in the combined system because they are deemed distinctive for regression and progression cases. Regression and progression difference images are opposite to each other in terms of intensity, therefore we expect the intensity distribution features to be able to distinguish between them without the need for more sophisticated features.

The k-NN classifier was employed in first stage of classification, and the LDA in the second stage. The k-NN classifier is a nonparametric classifier. According to the k-NN rule, the test sample is assigned the majority label of the nearest k training samples. The free parameter k has to be chosen experimentally ($k=15$ in our system). In this work, the fast implementation of the k-NN classifier by Arya and Mount[23] was used.

IV. EXPERIMENTS

IV.A. Experimental setup

Classification experiments in this study were performed using the leave-one-out cross validation procedure. Cross validation involves training a classifier n times, each time leaving out one of the n disjoint data subsets from training, and using only the omitted subset for validation. For leave-one-out cross validation, n equals the sample size. This technique guarantees the optimal use of the available data.

We divided the data into training and test sets on the basis of scans, not sections. Therefore, one to three image pairs of the same patient were set aside in each leave-one-out iteration. In this way, at no time did training and test sets contain samples originating from the same scan pairs. For each classification strategy the system performed 74 classifications. The final accuracy was computed from the outcomes of all classifications.

For strategy II, classification was performed as follows in each iteration of the leave-one-out procedure. In the first stage the entire training set was used to train the classifier. The training image pairs were relabeled from regression, stable, and progression to changed and stable in order to perform two-class classification. The second stage classification was applied only to those test samples that were classified as changed in the first stage. To train the second classifier, stable pairs were removed from the training set, thus, the classifier was trained only with regression and progression cases.

IV.B. Choice of system parameters

The system performance was evaluated with different choices for distance measures and classifiers. We compared the following distance measures: Minkowski distances of orders of 1 and 2, χ^2 statistics, and Jeffrey divergence (see, for example, Ref. 22 for details on these distance measures). Four different classifiers were considered to use with our

TABLE I. The performances of two classification strategies and two observers, measured on the same data set by accuracy and by weighted kappa κ. Here κ measures the agreement of a rater with the reference standard.

	Strategy I	Strategy II	Observer I	Observer II
Accuracy	0.761	0.795	0.785	0.820
κ	0.611	0.683	0.729	0.740

system: the linear discriminant analysis (LDA), the k nearest neighbor classifier (k-NN), and support vector machines with linear and exponential kernels (see, for example, Ref. 24 for the description of different classifiers). Due to the large number of features in the first stage of strategy II, the principal component analysis (PCA) was considered as the means of dimensionality reduction prior to classification. We used the same experimental data and the leave-one-out setup for comparing the classification systems with different parameters.

In both classification strategies, feature vectors were normalized to zero mean and unit variance prior to classification. Normalization parameters were estimated in the training data and used on the feature vectors of the test data.

IV.C. Observer study

In order to compare the performances of the two classification strategies to that of radiologists, an observer study was conducted. Observers were presented the same set of 205 registered pairs. Prior to that, pairs were randomly shuffled, so that pairs of sections from the same scans (one to three pairs) did not necessarily follow each other. Additionally, before presenting each pair for a side-by-side comparison, the side of the display on which the baseline scan was projected was randomly chosen. The observer was asked to classify the extent of disease in the image on the right-hand side compared to the image on the left-hand side. There were three classification categories—decrease (disease extent reduction $>2\%$), stable (any change in the disease extent $\leqslant 2\%$), and increase (disease extent increase $>2\%$). Both observers were chest radiologists in training. They were not involved in setting the reference standard for the data in this study.

V. RESULTS

The classification performances of the system and the observers were estimated by means of accuracy and weighted kappa statistics. In the explanation of the results we use the term "rater" to indicate either a computer system or a human observer.

The LDA was the most accurate classifier for strategy I. The best classification accuracy for strategy II was obtained with the Minkowski distance of order of 1 (also known as the city block distance) and k-NN classifier ($k=15$) in first stage of classification and the LDA in the second stage. In this work, the fast implementation of the k-NN classifier by Arya and Mount[23] was used. The classification accuracy in the first stage of strategy II benefited from applying PCA. PCA retained 99% of variance in the feature vector and reduced the feature space to 48 components. In this section, the classification results are given only for the best performing system.

Classification accuracy was calculated as the fraction of correctly classified samples in the test data set. In Table I accuracies are shown for each system and observer. The kappa statistics is a popular measure of inter-rater agreement on ordinal or nominal scales in medical research.[25] We used the linearly weighted kappa statistics[26] which penalizes disagreement between stable and regression or progression less than disagreement between regression and progression. The kappa statistics was computed for each rater versus the reference standard agreement as well as for inter-rater agreement for all pairs of raters. The values of kappa are given in Tables I and III, respectively. From Table I it follows that the classification strategy II showed better accuracy and agreement with the reference standard than strategy I, and the accuracy of strategy II was close to that of human observers.

The corresponding contingency matrices are presented in Table II. In each matrix, rows represent the reference standard and columns represent a rater's opinion. Entries on a

TABLE II. Contingency matrices of the two classification strategies and the two observers versus the reference standard. Each row represents the reference standard. Each column represents a class obtained by the system or observer. Class names are abbreviated: R for regression, S for stable, and P for progression.

		Strategy I				Strategy II	
True class	R	S	P	True class	R	S	P
R	27	24	0	R	33	18	0
S	1	101	3	S	5	97	3
P	2	19	28	P	2	14	33
True class		Observer I		True class		Observer II	
	R	S	P		R	S	P
R	47	3	1	R	38	13	0
S	15	66	24	S	8	92	5
P	1	0	48	P	1	10	38

TABLE III. Inter-rater agreement measured by weighted kappa.

	Strategy I	Strategy II	Observer I
Strategy II	0.747		
Observer I	0.444	0.505	
Observer II	0.638	0.653	0.640

matrix diagonal show numbers of pairs correctly classified in each class. The contingency matrices show different tendencies in misclassification for different raters. For example, observer I seems more sensitive to small differences in the disease extent than the other raters. Observer I has the fewest misclassifications in regression and progression categories, but, at the same time, a lot of misclassifications of stable cases for cases with change. The bias of the computer strategies is different—both of them were inclined to erroneously classify pairs with change as stable. The values of kappa in Table III support this observation. Both classification strategies show moderate agreement with observer I, but good agreement with observer II. Observer II also shows good agreement with observer I, which possibly indicates that observer II did not have a clear classification tendency.

VI. DISCUSSION

Here we discuss the results in more detail.

Neither the classification strategies nor the observers had much difficulty in correctly classifying obvious changes. As mentioned in Sec. II B, the reference standard was initially given on a seven-point scale, with two extreme points indicating “massive decrease” and “massive increase,” respectively. All 28 pairs in these categories were correctly classified as either regression or progression by both observers. Strategy I correctly classified 25 pairs, making mistakes in three pairs belonging to the same patient. Strategy II made the same errors as strategy I and additionally misclassified one pair. An example of a correctly classified case with massive disease progression is given in Fig. 1(c). The three pairs misclassified by both strategies exhibited a type of abnormality that was unique in the collected data set. This means that the classification algorithms did not have an opportunity to train on similar patterns. One of these three pairs is shown in Fig. 2(a). The histogram image in this figure suggests that this pair might have been misclassified by strategy I even in the presence of similar patterns in the training data because the difference image intensity histogram does not reflect such a subtle change.

Among 37 pairs with “moderate” changes, which meant 10%–50% decrease or increase in disease extent, both strategy II and observer II made 6 errors, while observer I made no errors, and strategy I made 16 errors. All errors pertained to mislabeling a pair with change as a stable case. Only three image pairs with moderate change, belonging to two different patients, were misclassified by both strategies but correctly classified by both observers. Two of these three pairs are shown in Figs. 4(a) and 4(b). An example of a case with moderate disease regression correctly classified by all raters is given in Fig. 1(a). Some cases were correctly classified by one of the classification strategies but misclassified by observer II: two examples are shown in Figs 4(c) and 4(d).

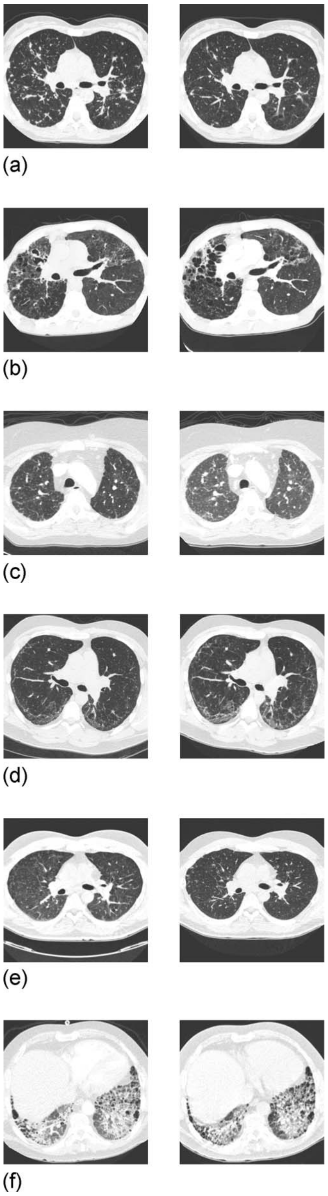

FIG. 4. Examples of ILD progression classification in follow-up images. Each example consists of a baseline CT section (left image) and a corresponding follow-up CT section (right image). Pairs (a) and (b): each pair exhibits moderate disease regression (between 10% and 50%) and was correctly classified by both human observers but misclassified by both classification strategies. Pairs (c) and (d): each pair exhibits moderate disease progression (between 10% and 50%) and was correctly classified by one or both classification strategies, but misclassified by one of the human observers. Pairs (e) and (f): each pair exhibits minor disease progression (between 2% and 10%), but was misclassified for disease regression by both classification strategies and one or both radiologists.

Most misclassifications occurred for cases from the “minor decrease” and “minor increase” categories (2%–10% change), both for the observers and the classification algorithms. Among 35 pairs with “minor” changes, half were misclassified by observer II, and more than half by both classification strategies. However, observer I made only five errors with these pairs. It is possible that the threshold of 2% used in our protocol lies within statistical uncertainty and cannot be used reliably. In the literature, thresholds of 10% (Ref. 27) and 5% (Ref. 12) have been used to differentiate between changed and stable diseases.

Very few errors were made by confusing regression and progression cases: both computer strategies and observer I made two errors, while observer II made one error. Only three image pairs were misclassified in this way by all raters jointly, two of them being from the same patient. The misclassified pairs always exhibited a minor change in disease extent. Two of the three pairs, taken from different patients, are shown in Figs. 4(e) and 4(f). They clearly represent difficult cases.

The most evident difference between the two strategies is the improvement of the recognition of change by strategy II compared to strategy I, as seen from Table II. Figures 2(b), 2(c), and 4(d) demonstrate image pairs with ILD progression that were classified correctly by strategy II but misclassified as stable by strategy I. The observed improvement could be attributed to a more appropriate set of features for discrimination between changed and stable image pairs employed in the first stage of strategy II. Additionally, it should be noted that the training set in the first stage of strategy II was balanced, with approximately the same number of samples in both changed and stable classes. At the same time, the training set in strategy I contained twice as many samples of the stable class as either of the other two classes. This might have negatively influenced the performance of strategy I.

The results obtained by our automated analysis were promising because it exhibited good agreement with the ground truth and accuracy close to that of human observers ($\kappa=0.683$ and an accuracy 79.5% for strategy II). The current concept of CAD does not require a computer performance to be better or equal to that of radiologists,[28] but it needs to be complementary to that of radiologists. It is interesting to note that agreement between observer I and each classification strategy was much lower than the agreement between observer II and each strategies (for strategy II, $\kappa=0.505$ and $\kappa=0.653$, respectively). Because the accuracy of strategy II was similar to that of observer I, it is likely that observer I would benefit more from the complementarity of the computer output. Conducting an observer study to show whether using our CAD scheme improves the performance of radiologists on this task is a topic for future research.

Several possible improvements in the system setup can be identified.

We used a single expert's annotations as the reference standard. As we have already mentioned in Sec. I, manual annotations are not considered reliable for training a texture classification analysis because of low rating agreement. Our task, however, required the estimation of overall disease extent, which should cause less intra- and interobserver variabilities between experienced chest radiologists. As opposed to labeling small regions of interest into multiple categories, assessing disease extent is part of the daily clinical routine for radiologists. In clinical studies on ILD prognosis and progression, a single expert's estimate is commonly referred to as the ground truth. For example, in Ref. 12, the visual assessment of the disease extent was used as one of three independent criteria to define a patient's disease progression, along with physiologic tests, such as total lung capacity and the resting oxygen saturation level. Combining an expert's estimates, or estimates obtained by consensus, with the results of physiologic tests may improve the reliability and consistency of our reference standard. Another improvement of the training process would be the enlargement of the training data set, so that different abnormal patterns are sufficiently represented. Evaluating the system on a different data set, preferably, from another institution, is also a topic for future research.

In order to train the classification system on well-represented classes, we performed a balancing procedure described in Sec. II B. In a number of randomly selected pairs the baseline and follow-up images were swapped and the class labels of such pairs were changed to opposite labels, accordingly. But we do not know how legitimate such a procedure is from a clinical point of view, in other words, whether an artificially created progression could be distinguished by a radiologist from the sequence of images representing a real disease progression. As a topic for future research, an observer study could be conducted that investigates its validity.

There is an inherent limitation of the system performance related to the generalizing ability of dissimilarity-based features. Dissimilarities computed over the whole lung fields are likely to neglect some small local changes that would probably be revealed in dissimilarities over smaller areas. A recent study[10] showed a good performance in classification of small VOIs by means of dissimilarities. In that study, a more sophisticated dissimilarity measure than in our study and a different way of histogram binning were used. Experimenting with measures of dissimilarity between histograms, other than those mentioned in Sec. III E, could be beneficial for our system as well. Application of our system to regions smaller than a CT section is also possible provided the ground truth for smaller areas is available.

It is computationally efficient to train and apply automated analysis to two-dimensional CT sections. Modern clinical practice requires, however, the analysis of 3D volumes. In our system, image preprocessing (intrapatient registration and lung segmentation) is already done for full CT scans. The computation of features and classification strategies presented in this paper could be easily extended to 3D. An alternative approach would be to apply our system section by section to the 3D scan and subsequently fuse the classification outcomes of individual sections into a decision about the whole lung volume or its part. This approach might be preferable to the direct 3D analysis because the generalizing quality of dissimilarity features is likely to be enhanced in 3D which will make the system less sensitive to small changes.

VII. CONCLUSIONS

We have developed a classification system that performs estimation of ILD progression in axial sections extracted from serial thoracic CT scans. To achieve this, our system comprises nonrigid intrapatient image registration, multiatlas lung segmentation, texture feature extraction, and computation and classification of dissimilarities. The system employing classification strategy II showed good agreement with the reference standard, and its accuracy approached that of two radiologists.

ACKNOWLEDGMENTS

This research was supported by the Dutch Technology Foundation STW, Applied Science Division of NWO and the technology programme of the Dutch Ministry of Economic Affairs.

[a] Present address: CSIRO Mathematical and Information Sciences, 2113 Sydney, Australia; electronic mail: yulia.arzhaeva@csiro.au

[1] British Thoracic Society, "BTS guidelines on the diagnosis, assessment and treatment of diffuse parenchymal lung disease in adults," Thorax **54**, S24–S30 (1999).

[2] G. Raghu and K. K. Brown, "Interstitial lung disease: clinical evaluation and keys to an accurate diagnosis," Clin. Chest Med. **25**, 409–419 (2004).

[3] E. A. Kazerooni, "High-resolution CT of the lungs," AJR, Am. J. Roentgenol. **177**(3), 501–519 (2001).

[4] C. Schaefer-Prokop, M. Prokop, D. Fleischmann, and C. Herold, "High-resolution CT of diffuse interstitial lung disease: Key findings in common disorders," Eur. Radiol. **11**, 373–392 (2001).

[5] K. Murphy, M. Prokop, C. M. Schaefer-Prokop, H. Gietema, G. D. Nossent, B. van Ginneken, J. P. W. Pluim, and Y. Arzhaeva, "Improved efficiency of assessment of interstitial lung disease progression in CT of the chest by visualisation of automatically-registered image pairs," Radiological Society of North America, Annual Meeting, 2008 (unpublished), Vol. 94.

[6] I. C. Sluimer, M. Prokop, I. Hartmann, and B. van Ginneken, "Automated classification of hyperlucency, fibrosis, ground glass, solid and focal lesions in high resolution CT of the lung," Med. Phys. **33**(7), 2610–2620 (2006).

[7] H. J. Kim, G. Li, D. Gjertson, R. Elashoff, S. K. Shah, R. Ochs, F. Vasunilashorn, F. Abtin, M. S. Brown, and J. G. Goldin, "Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study," Acad. Radiol. **15**(8), 1004–1016 (2008).

[8] I. C. Sluimer, A. M. R. Schilham, M. Prokop, and B. van Ginneken, "Computer analysis of computed tomography scans of the lung: A survey," IEEE Trans. Med. Imaging **25**(4), 385–405 (2006).

[9] Y. Xu, E. J. R. van Beek, Y. Hwanjo, J. Guo, G. McLennan, and E. A. Hoffman, "Computer-aided classification of interstitial lung diseases via MDCT: 3D adaptive multiple feature method (3D AMFM)," Acad. Radiol. **13**(8), 969–978 (2006).

[10] V. A. Zavaletta, B. J. Bartholmai, and R. A. Robb, "High resolution multidetector CT-aided tissue analysis and quantification of lung fibrosis," Acad. Radiol. **14**(7), 772–787 (2007).

[11] H. Sumikawa, T. Johkoh, S. Yamamoto, K. Takahei, T. Ueguchi, Y. Ogata, M. Matsumoto, Y. Fujita, J. Natsag, A. Inoue, M. Tsubamoto, N. Mihara, O. Honda, N. Tomiyama, S. Hamada, and H. Nakamura, "Quantitative analysis for computed tomography findings of various diffuse lung diseases using volume histogram analysis," J. Comput. Assist. Tomogr. **30**(2), 244–249 (2006).

[12] A. C. Best, J. Meng, A. M. Lynch, C. M. Bozic, D. Miller, G. K. Grunwald, and D. A. Lynch, "Idiopathic pulmonary fibrosis: Physiologic tests, quantitative CT indexes, and CT visual scores as predictors of mortality," Radiology **246**(3), 935–940 (2008).

[13] I. C. Sluimer, P. F. van Waes, M. A. Viergever, and B. van Ginneken, "Computer-aided diagnosis in high-resolution CT of the lungs," Med. Phys. **30**(12), 3081–3090 (2003).

[14] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," IEEE Trans. Med. Imaging **18**(8), 712–721 (1999).

[15] P. Thévenaz, T. Blu, and M. Unser, *Handbook of Medical Imaging, Processing and Analysis* (Academic, San Diego, 2000), pp. 393–420.

[16] S. Klein, M. Staring, and J. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines," IEEE Trans. Image Process. **16**, 2879–2890 (2007).

[17] S. Hu, E. Hoffman, and J. Reinhardt, "Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images," IEEE Trans. Med. Imaging **20**, 490–498 (2001).

[18] I. Sluimer, M. Prokop, and B. van Ginneken, "Towards automated segmentation of the pathological lung in CT," IEEE Trans. Med. Imaging **24**(8), 1025–1038 (2005).

[19] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer Jr., "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," Neuroimage **21**(4), 1428–1442 (2004).

[20] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," Neuroimage **33**(1), 115–126 (2006).

[21] D. A. Lynch, W. D. Travis, N. L. Müller, J. R. Galvin, D. M. Hansell, P. A. Grenier, and T. King Jr., "Idiopathic interstitial pneumonias: CT features," Radiology **236**, 10–21 (2005).

[22] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," Int. J. Comput. Vis. **40**(2), 99–121 (2000).

[23] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," J. ACM **45**(6), 891–923 (1998).

[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (Wiley, New York, 2001).

[25] D. G. Altman, *Practical Statistics for Medical Research* (Chapman and Hall, London/CRC, Boca Raton, 1991).

[26] J. Cohen, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," Psychol. Bull. **70**, 213–220 (1968).

[27] N. J. Screaton, M. P. Hiorns, K. Lee, T. Franquet, T. Johkoh, K. Fujimoto, K. Ichikado, T. V. Colby, and N. L. Müller, "Serial high resolution CT in non-specific interstitial pneumonia: prognostic value of the initial pattern," Clin. Radiol. **60**(1), 96–104 (2005).

[28] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," Comput. Med. Imaging Graph. **31**(4–5), 198–211 (2007).