Toward automatic regional analysis of pulmonary function using inspiration and expiration thoracic CT

Keelin Murphy[a] and Josien P. W. Pluim
Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands

Eva M. van Rikxoort
Department of Radiology, Radboud University Nijmegen Medical Centre, 6525 GA Nijmegen, The Netherlands

Pim A. de Jong
Department of Radiology, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands

Bartjan de Hoop
Department of Radiology, Rijnstate Hospital, 6815 AD Arnhem, The Netherlands

Hester A. Gietema and Onno Mets
Department of Radiology, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands

Marleen de Bruijne
Department of Computer Science, University of Copenhagen, DK 2100 Copenhagen, Denmark and Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC—University Medical Center, 3015 CE Rotterdam, The Netherlands

Pechin Lo
Department of Computer Science, University of Copenhagen, DK 2100 Copenhagen, Denmark

Mathias Prokop and Bram van Ginneken
Department of Radiology, Radboud University Nijmegen Medical Centre, 6525 GA Nijmegen, The Netherlands

(Received 22 August 2011; revised 12 December 2011; accepted for publication 4 February 2012; published 1 March 2012)

Purpose: To analyze pulmonary function using a fully automatic technique which processes pairs of thoracic CT scans acquired at breath-hold inspiration and expiration, respectively. The following research objectives are identified to: (a) describe and systematically analyze the processing pipeline and its results; (b) verify that the quantitative, regional ventilation measurements acquired through CT are meaningful for pulmonary function analysis; (c) identify the most effective of the calculated measurements in predicting pulmonary function; and (d) demonstrate the potential of the system to deliver clinically important information not available through conventional spirometry.

Methods: A pipeline of automatic segmentation and registration techniques is presented and demonstrated on a database of 216 subjects well distributed over the various stages of COPD (chronic obstructive pulmonary disorder). Lungs, fissures, airways, lobes, and vessels are automatically segmented in both scans and the expiration scan is registered with the inspiration scan using a fully automatic nonrigid registration algorithm. Segmentations and registrations are examined and scored by expert observers to analyze the accuracy of the automatic methods. Quantitative measures representing ventilation are computed at every image voxel and analyzed to provide information about pulmonary function, both globally and on a regional basis. These CT derived measurements are correlated with results from spirometry tests and used as features in a kNN classifier to assign COPD global initiative for obstructive lung disease (GOLD) stage.

Results: The steps of anatomical segmentation (of lungs, lobes, and vessels) and registration in the workflow were shown to perform very well on an individual basis. All CT-derived measures were found to have good correlation with spirometry results, with several having correlation coefficients, r, in the range of 0.85–0.90. The best performing kNN classifier succeeded in classifying 67% of subjects into the correct COPD GOLD stage, with a further 29% assigned to a class neighboring the correct one.

Conclusions: Pulmonary function information can be obtained from thoracic CT scans using the automatic pipeline described in this work. This preliminary demonstration of the system already highlights a number of points of clinical importance such as the fact that an inspiration scan alone is not optimal for predicting pulmonary function. It also permits measurement of ventilation on a per lobe basis which reveals, for example, that the condition of the lower lobes contributes most to

the pulmonary function of the subject. It is expected that this type of regional analysis will be instrumental in advancing the understanding of multiple pulmonary diseases in the future. © *2012 American Association of Physicists in Medicine*. [http://dx.doi.org/10.1118/1.3687891]

Key words: CT, automatic, ventilation, airflow, pulmonary, registration, segmentation, functional, spirometry, COPD

I. INTRODUCTION

Acquiring a thoracic CT scan is one of the most common and effective ways to examine the condition of the in vivo lung. A breath-hold CT scan provides detailed information about the site and severity of many possible pathologies. It has been observed that in most cases a low-dose protocol provides an image with acceptable diagnostic quality while limiting the radiation exposure of the subject.[1–4] Although inspiration breath-hold is the most frequently acquired type of scan, it is increasingly recognized that an additional breath-hold expiration scan adds considerably to the diagnostic information obtained and can be acquired with relatively little increase to the overall radiation dose.[5–7]

The current gold standard in analyzing lung function is the pulmonary function test (spirometry), which provides only global information about the levels of inspiration and expiration of the patient. Methods of regional functional analysis using CT imaging are discussed in van Beek *et al.*[8] which require the use of xenon gas or other contrast agents. Due to expense or impracticality such methods have not become widely used. Comparing conventional breath-hold inspiration and expiration CT scans from a subject can, however, provide important information about ventilation and airflow. In particular, such analysis allows for the determination of regional severity of airflow limitation, for example, on a per lobe basis. However, such a comparison is difficult and time consuming to carry out visually, particularly since anatomical structures are not well aligned between the scans. Automatic image registration techniques may be used to align the pulmonary structures correctly and furthermore to enable an automatic, quantitative, regional comparison of lung densities.

The vast majority of literature on processing chest CT scans deals with single scan processing,[9,10] while some authors have also combined inspiration and expiration scan processing, calculating, and comparing features of the individual scans such as lung volumes.[11–13] However, works which involve a full 3D registration of inspiration and expiration data prior to density analysis are relatively few and recent, possibly since this registration task is extremely difficult due to the large deformations involved. In a recent public challenge to register 30 pairs of thoracic CT scans,[14,15] the pairs involving full inspiration and full expiration data proved to be the most difficult among all those supplied.

Reinhardt *et al.*[16] compare registration based measurements (Jacobian values) from respiratory gated CT images to ventilation measures acquired using xenon CT imaging in five sheep. Guerrero *et al.*[17] register phases from 4D-CT datasets of three subjects to obtain quantitative ventilation measurements. These measurements are compared with corresponding lung volume changes. Registration of 4D-CT is used to create ventilation images in Ref. 34 using 12 subjects and comparing two different registration techniques as well as two possible ventilation measurements (intensity measurements and Jacobian values). Intermethod variability only was studied without reference to any particular gold standard. Comparison of Jacobian values with intensity differences is also carried out by Ding *et al.*,[18] using three sheep with xenon CT measurements as the gold standard. In Refs. 35 and 33, the authors use ventilation images extracted from registered 4D-CT to identify high functioning lung regions in radiotherapy planning (15 subjects) and to identify low functioning regions in emphysema patients (12 subjects). Castillo *et al.*[19] register 4D-CT images from seven subjects to produce ventilation images which are compared with SPECT (single photon emission CT) imaging as well as with global volume change measurements. Two types of Jacobian measurement as well as a density change calculation are experimented with in the creation of ventilation images.

In this work, we present an automatic method for regional comparison of parenchymal density at full inspiration and full expiration. In contrast with previous works mentioned above, a much larger database is used for testing and the conventional gold standard of spirometry test results is available to us as a reference standard. In addition, each step of the process is analyzed and evaluated to determine at which points errors are likely to occur, and how this may impact the final results. A series of different ventilation measurements is investigated for comparison purposes. The method comprises a pipeline of individual processing steps including the segmentation of lungs, fissures, airways, lobes, and vessels, registration of the two images and computation of quantitative measures relating to ventilation at each image voxel. Such a system, in contrast with pulmonary function testing alone, has the potential to provide rich detail about which pathologies are present and which regions of the lungs are more or less affected by airflow limitation. In the future, this type of information may prove to be key in unravelling the mechanisms behind major lung diseases such as chronic obstructive pulmonary disease (COPD), asthma, sarcoidosis, and cystic fibrosis.

As an example of how the system may be utilized, it is applied to a database of 216 subjects with COPD at varying global initiative for obstructive lung disease (GOLD) (Ref. 21) stages. The aims of this work are as follows:

- To analyze the workflow leading to the calculation of the quantitative measurements in detail. Automatic segmentations and registrations are each examined and scored by radiology experts to identify any weak points in the pipeline.
- To validate the CT-derived ventilation measurements. All measurements are correlated with the current gold standard

of spirometry results and are furthermore used in a kNN classifier to predict GOLD stages. In this way, we aim to verify whether the measurements truly represent airflow limitation.
- To determine which measurements are most useful in predicting pulmonary function. This will imply suggestions of how best to tune the system in the future.
- To highlight the advantages of this method in comparison with conventional pulmonary function testing and use results from these preliminary experiments to reveal clinically relevant information.

Section II gives an introduction to COPD, while Sec. III describes the database of COPD subjects used to demonstrate the system. In Sec. IV, the various components of the CT-based ventilation measurement system are described in detail and the experiments performed are detailed in Sec. V. The results of the experiments are provided in Sec. VI while Sec. VII provides an overall discussion of the system and results described.

II. CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD)

In this work, a database of COPD subjects will be used to demonstrate the automatic method being described. We therefore provide a brief introduction to COPD in this section.

COPD is expected to be the fourth leading cause of death globally by 2030.[20] The illness is characterized by airflow limitation which is not fully reversible and diagnosis is confirmed by means of spirometry (pulmonary function testing).[21] During pulmonary function tests subjects are instructed to exhale fully into a mouthpiece and various measurements are made, including "forced expiratory volume in 1 second" (FEV1) and "forced vital capacity" (FVC). FEV1 measures how much air volume is released in the first second of expiration, and FVC determines the entire volume exhaled. A value known as "FEV1-predicted" defines the FEV1 in the average population, taking into account details such as gender, height, weight, age, and race. This value is determined from look-up tables.[22]

The global initiative for obstructive lung disease (GOLD) defines four levels of severity according to the spirometry results (GOLD stages 1–4),[21] and the status of subjects without COPD is here referred to as "stage 0." Table I illustrates how a COPD GOLD stage is diagnosed based on these measurements. FEV1 as a percentage of FVC (FEV1/FVC%) determines the presence or absence of COPD, while FEV1 as a percentage of FEV1-predicted (FEV1%predicted) is used to establish the severity of the condition. There may be several underlying causes of the airflow obstruction including emphysema and/or small airway disease. In thoracic CT, these pathologies are most notably characterized by parenchymal destruction and air-trapping.

TABLE I. The criteria for determination of COPD class by pulmonary function testing

COPD class	0 (no COPD)	1 (Mild)	2 (Moderate)	3 (Severe)	4 (Very Severe)
Measurement					
FEV1/FVC%	≥70	<70	<70	<70	<70
FEV1%predicted		≥80	≥50, <80	≥30, <50	<30

III. MATERIALS

To illustrate the utility of the system, a database was constructed consisting of subjects with various stages of COPD severity. When constructing the database it was attempted to maximize the number of subjects while keeping a good balance of the various GOLD stages. Subjects with stage 4 COPD were particularly difficult to find however, therefore, there are somewhat fewer of these in the final selection. The final database consisted of 216 subjects with 49 of these at each of GOLD stages 0 and 1, 50 at GOLD stage 2, 42 at GOLD stage 3, and 26 at GOLD stage 4.

For each of the 216 subjects, a breath-hold scan at both inspiration and expiration (taken on the same day) was available. The subjects consisted of 188 males and 28 females, aged 52–79 years, with an average age of 61 years. The data were drawn from the Nelson screening trial[23] (172 subjects) and from clinical practice at the University Medical Center in Utrecht, The Netherlands (44 subjects). The signal-to-noise ratio in the scans was relatively low in many cases due to the fact that they were acquired using a low-dose (30 mAs) or ultralow-dose (20 mAs) protocol. Of the 432 scans, the beam current, c, was $\leq$20 mAs for 203 of these, while a further 176 scans had 20 mAs $< c \leq$ 30 mAs. All scans had 512×512 pixels per slice with the number of slices varying from 321 up to 677 (average 463 slices). Because the smallest field of view was chosen which included the outer rib margins at the widest dimensions of the thorax, voxel spacing varied from 0.45 mm up to 0.89 mm with an average of 0.69 mm for each direction. Slice thickness varied from 0.45 to 0.7 mm with an average thickness of 0.68 mm.

IV. METHODS

In this section, the pipeline of processing involved in order to obtain ventilation measurements from CT scan pairs is described in detail. Briefly, following some initial preprocessing, the lungs were automatically segmented. These segmentations were used in the registration of the inspiration image with the expiration image for each subject as well as in the segmentation of further anatomical structures including lobes and vessels. Using the original images, the registered images, and the segmentation information, a variety of ventilation measurements was computed. Figure 1 depicts an overview of the system pipeline. All processing was carried

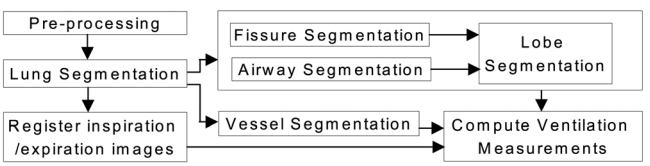

FIG. 1. Flowchart depicting the pipeline of processing for CT scans.

out on a desktop PC with an Intel Core 2 Quad CPU, 2.4 GHz, and 8 GB of RAM.

IV.A. Preprocessing

Processing of CT scans can be a memory-intensive task, particularly for high-resolution data. For this reason, some of the segmentation tasks described in this work use subsampled data. To subsample the data, the image size was reduced by block averaging to 256×256 voxels in the X-Y plane with the number of slices reduced such that the data were isotropically sampled. Linear interpolation was used to determine gray-values between voxel locations. This strategy does not attempt to apply a consistent image spacing for all images but rather aims to retain the best resolution possible for each individual image. In this way, we hope to achieve the highest chance of success for each image in subsequent processing tasks.

In order to reduce memory consumption in processes where full resolution data are preferred, the scan size was reduced by excluding image regions outside the lungs (after lung segmentation has taken place). A bounding box around the segmented lungs was constructed, with a margin of 5 voxels on each side. Data outside this bounding box were discarded. The resulting smaller image is referred to in this work as the bounded image.

IV.B. Lung segmentation

Lung segmentation is carried out on the subsampled data using the method of van Rikxoort *et al.*[24] This is a hybrid method of lung segmentation specifically designed to overcome segmentation errors typical in scans exhibiting pathology. The process begins with a conventional approach to segmenting the lungs using region-growing and morphological smoothing. An automatic error-checking process then commences, whereby measures such as the volumes of the segmented left and right lungs, the ratio between these volumes, and the shape consistency of the lung borders are checked to determine whether the segmentation is likely to be erroneous. If the measures indicate that the segmentation was successful no further action is taken. However, scans where these measures appear abnormal are reprocessed using a more computationally expensive multi-atlas segmentation algorithm. Of the database of 432 scans used in this work, 30 of them were segmented using this multi-atlas method.

IV.C. Fissure segmentation

Fissures are segmented as part of our workflow since they are required in the lobe segmentation (see Sec. IV E). The method of extracting the fissures is based on a voxel classification scheme using a supervised classifier. This is described in detail by van Rikxoort *et al.*[25]

IV.D. Airway segmentation

Where no fissure information is available (because of subject anatomy or scan quality), the airways are used to guide the lobe segmentation (see Sec. IV E). The airways are extracted using a region-growing approach with adaptive thresholding, as described by van Ginneken *et al.*[26]

IV.E. Lobe segmentation

The pulmonary lobes are segmented from the subsampled data using a multi-atlas based approach from van Rikxoort *et al.*[27] This approach depends on other automatic segmentations including segmentations of the lungs, the fissures, and the airways as described above. If fissures are segmented successfully, then these are used to define the lobar borders. However, in some cases, fissures are fully or partially absent, or very difficult to segment due to image quality. Airway segmentations are used to approximate the lobar borders where no fissure information is available. Five atlas images with known lobe segmentations are provided for each of the left and right lungs. The anatomical information provided by the fissure/airway segmentations is used to determine the most similar atlas, which is then registered to the scan. The lobe segmentation from the registered atlas image is then propagated to the scan to provide the final lobe segmentation.

IV.F. Vessel segmentation

In order to exclude vessels from ventilation measure calculations (in which only lung parenchyma is of interest), a vessel segmentation is performed. The vessels were segmented using the algorithm of Lo *et al.*,[28] which selects voxels as vessel or nonvessel based on features of brightness, contrast, and tubeness. Since many of the scans used a low-dose or an ultralow dose protocol, the best results were obtained by using the full resolution scan data in the bounded image and applying a noise filter[29] prior to the segmentation. A slight oversegmentation was preferred to an undersegmentation since the purpose was to ensure that vessel structures were not included in calculations related to parenchymal tissue. Therefore, some postprocessing was applied to the segmentation to ensure that as many vessels as possible were captured. First, the vessel segmentation was dilated with a spherical kernel of radius 1 voxel. In addition, a conservative thresholding was applied to the original image to segment the brightest voxels. The final segmentation comprised the union of the dilated vessel segmentation and the thresholded bright voxels. The threshold t was set at $t = \mathrm{CT}_{\mathrm{min}} + [(\mathrm{CT}_{\mathrm{max}} - \mathrm{CT}_{\mathrm{min}}) \times 0.5]$ where $\mathrm{CT}_{\mathrm{min}}$ and $\mathrm{CT}_{\mathrm{max}}$ represent the minimum and maximum intensity values within the lung volume. In practice, to avoid $\mathrm{CT}_{\mathrm{min}}$ and $\mathrm{CT}_{\mathrm{max}}$ being influenced by noise or minor segmentation errors, $\mathrm{CT}_{\mathrm{min}}$ is given the value of the first centile and $\mathrm{CT}_{\mathrm{max}}$ the value of the 99th centile.

IV.G. Registration

All registrations were performed using *elastix* version 4.3 (Refs. 30 and 31). In each pair, the expiration scan was considered as the "fixed" (target) image, while the inspiration scan was the "moving" (source) image. In other words, the inspiration scan was deformed to match the expiration scan. Registration of a scan pair was carried out using the bounded

images in three steps as described by Staring *et al.*[32] Briefly, these steps are as follows:

1. Affine registration using all visible anatomy in the bounded images, including structures outside the lungs. This gives an initial coarse alignment.
2. Nonrigid registration using preprocessed images where locations more than 2 voxels outside the lung boundary had been assigned the value 0. By removing structural information outside the lungs, the registration is concentrated only on aligning the lung volumes. The lung boundaries themselves are retained with a small margin around them to ensure that they are accurately aligned. The coarse alignment determined in the previous step is used as an initial transformation.
3. Nonrigid registration using preprocessed images as described in the previous step and calculating the cost function only on voxels which are within the lung volume of the fixed image. This use of the fixed lung mask improves the final alignment of small structures within the lungs. As before, the alignment determined in the previous step is used as an initial transformation.

A multiresolution strategy was used in all registrations with normalized correlation as a similarity measure and adaptive stochastic gradient descent as the optimization method.

IV.H. Local quantification of ventilation

Previous attempts to calculate ventilation information from CT scans[10,16–19,33–35] have used several different ventilation measures with varying results. Castillo *et al.*[19] found that density based measures have higher correlation with SPECT imaging than Jacobian measures, while Ding *et al.*[18] obtain better correlation with xenon-CT by using Jacobian measures. Since it has not been conclusively determined which measures are the most useful, a range of automatic quantitative measurements were calculated in this work. These included more conventional thresholding measures, on the original inspiration and expiration scans, which have previously been used in quantifying emphysema and air-trapping[36,37] as well as measures which use the registered inspiration and expiration data to evaluate the change in the lung parenchyma on a per voxel basis. Each measure was calculated for a total of eight regions as follows: (1) both lungs together (i.e., over total lung volume), (2)–(3) left and right lung, (4)–(8) each of the five lobes.

Those measures using the registered scan pairs compare CT values on a voxel-by-voxel basis. Voxels segmented as vessel were excluded from the calculations. These per voxel measures were then averaged (or the median was taken) over each of the eight regions defined above. Both the average and the median of the regional values were calculated in order to establish whether one gave substantially better results than the other.

Table II gives a complete list of the quantitative measures used along with explanations of how they were calculated and references where relevant. Each measure is labeled according to what measure is calculated and over which region (see region numbers provided above). Most of these measures were suggested in previous literature. In addition, we calculate subtraction values and ratio values from voxels in the registered scans since it seems intuitive that these would give an insight into the flow of air at each location.

The ventilation values v were calculated per voxel, using the registered inspiration and expiration scans, according to the following formula:

$$v = \frac{1000(\mathrm{HU}_{\mathrm{INSP}} - \mathrm{HU}_{\mathrm{EXP}})}{\mathrm{HU}_{\mathrm{EXP}}(\mathrm{HU}_{\mathrm{INSP}} + 1000)}, \quad (1)$$

where $\mathrm{HU}_{\mathrm{INSP}}$ and $\mathrm{HU}_{\mathrm{EXP}}$ are the CT densities in Hounsfield units at inspiration and expiration, respectively.[38] This measure reflects the percentage change in the amount of air at a location between full expiration and full inspiration. It is based on the theory that CT density change between expiration and inspiration is not linearly related to the amount of incoming air,[38] and therefore, methods such as subtracting HU values between expiration and inspiration are not appropriate.

Furthermore, ventilation measures using blood mass correction (BMC) are also calculated as was done by Guerrero.[17] The theory of BMC is based on the assumption that the total mass of the lungs is not changed by the actions of

TABLE II. The quantitative measures calculated automatically from the processed CT scans. The eight regions referred to in the labels are (1) both lungs together, (2)–(3) left and right lungs, (4)–(8) each of the five lobes (left upper, right upper, right middle, left lower, and right lower).

Name	Description	Labels (eight regions)
Perc850Exp	The percentage of volume with CT values below −850 Hounsfield units on the breath-hold expiration scan (Ref. 36).	A1–A8
Perc910Insp	The percentage of volume with CT values below −910 Hounsfield units on the breath-ho ld inspiration scan (Ref. 37).	B1–B8
Perc950Insp	The percentage of volume with CT values below −950 Hounsfield units on the breath-ho ld inspiration scan (Ref. 37).	C1–C8
AvgSubtr	The average of values obtained by subtracting the CT value at inspiration from the value at expiration (per voxel).	D1–D8
AvgRatio	The average of values obtained by taking the ratio of the CT value at inspiration to the value at expiration (per voxel).	E1–E8
AvgVentil	The average of ventilation values, calculated (per voxel) as described in Sec. IV H (Ref. 17).	F1–F8
AvgVentilBMC	The average of ventilation values with blood mass correction, calculated (per voxel) as described in Sec. IV H (Ref. 17).	G1–G8
MedSubtr	The median of values obtained by subtracting CT value at inspiration from the value at expiration (per voxel).	H1–H8
MedRatio	The median of values obtained by taking the ratio of the CT value at inspiration to the value at expiration (per voxel).	I1–I8
MedVentil	The median of ventilation values, calculated (per voxel) as described in Sec. IV H (Ref. 17).	J1–J8
MedVentilBMC	The median of ventilation values with blood mass correction, calculated (per voxel) as described in Sec. IV H (Ref. 17).	K1–K8

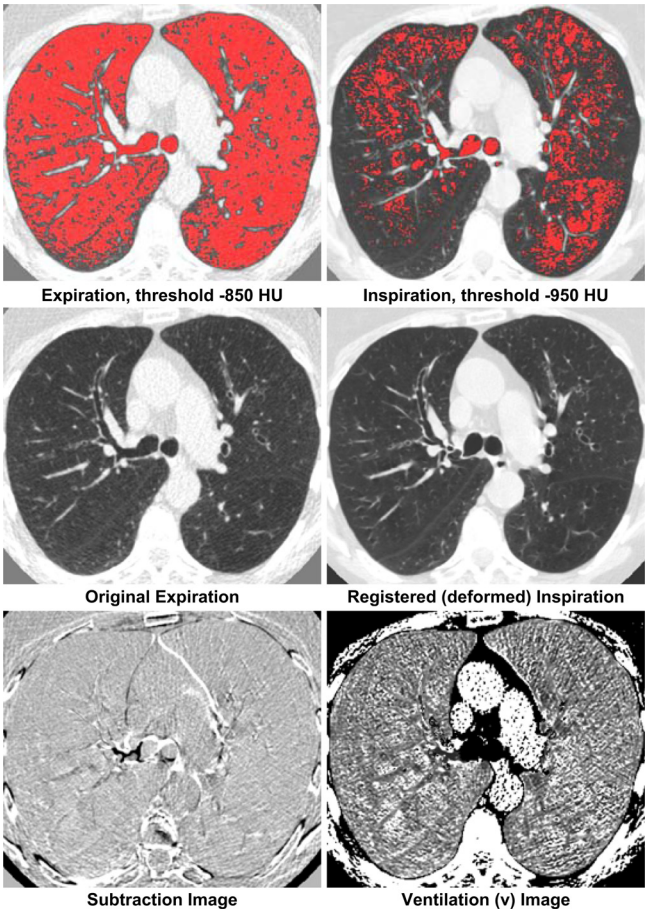

FIG. 2. Example images from a stage 4 COPD subject showing some of the features described in Table II. Note that only regions within the lung segmentation are considered during feature calculation. Top row: The expiration scan thresholded at −850 HU (features A1–A8, perc850Exp) and the corresponding inspiration scan (at approximately the same anatomical location) thresholded at −950 HU (features C1–C8, perc950Insp). Middle row: The original expiration image and the registered inspiration image which has been deformed to match it. These registered images are used to calculate all features D–K as described in Table II. Bottom row: Left: The subtraction image, used to calculate features D (avgSubtr) and H (medSubtr). Right: The ventilation image, used to calculate features F (avgVentil), G (avgVentilBMC), J (medVentil), and K (medVentilBMC).

inhaling and exhaling air. Calculating the masses according to the CT intensity values, however, it can be demonstrated that a minor change in mass occurs. This change in mass is due to alterations in blood-flow in the lungs during the breathing process.[39] Correcting for this involves calculating the fractional discrepancy f in mass between inspiration and expiration for each lung individually and then updating CT values on the inspiration scan, according to the following formula.[17]

$$\mathrm{HU}_{\mathrm{INSP-NEW}} = \mathrm{HU}_{\mathrm{INSP}} - 1000 f \left(1 + \frac{\mathrm{HU}_{\mathrm{INSP}}}{1000}\right). \quad (2)$$

The ventilation values may then be calculated using the corrected inspiration image and the expiration image as before.

Sample images from a stage 4 COPD subject are shown in Fig. 2. These images provide a visual interpretation of some of the features described in Table II.

V. EXPERIMENTS

V.A. Visual scoring

All scan pairs were processed according to the pipeline depicted in Fig. 1 and described in Sec. IV. Segmentations of the lungs, lobes, and vessels and scan pair registrations were all scored for accuracy by radiology experts. Each segmentation or registration was scored by a single observer from a group of four. The four observers were all radiology residents with a special interest in chest CT and between 1 and 10 years of experience. The options when scoring were as follows: "Excellent," "Good," "Moderate," and "Poor." These scores were provided for the segmentation of the left and right lungs, the segmentation of each of the five lobes, the segmentation of vessels in the left and right lungs, the registration of vessels in the left and right lungs, and the registration of fissures in the left and right lungs. The complete scoring protocol is given in appendix A. Although scores were provided at a regional (e.g. per lung, per lobe) level, and also at a per structure level for registration (fissures, vessels), for the purposes of this study these "subscores" were

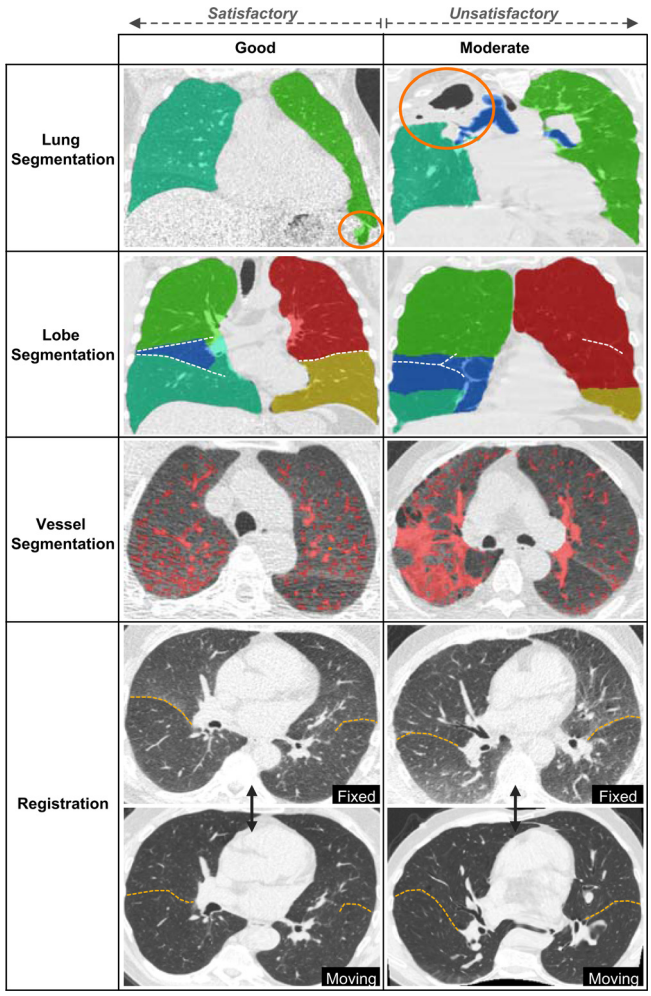

FIG. 3. Examples of scans scoring Good and Moderate for each type of segmentation and registration. Scans where scores were Good or better were retained (satisfactory), while those where scores were Moderate or worse were excluded (unsatisfactory). For a detailed description, see Sec. V A.

combined and the overall score for a segmentation/registration was given by the worst of the individual subscores.

A segmentation or registration that was scored as Excellent or Good was deemed to be satisfactory for the purposes of this study. Scan pairs where the segmentation or registration was not satisfactory were excluded. Figure 3 shows examples of scans in the Good and Moderate categories for each of the segmentation and registration procedures. These examples provide an indication of how the boundary between satisfactory and unsatisfactory is defined. In the lung segmentation examples, the Good example shows a slight leak into the bowel (circled), which persisted for only a few slices, while the rest of the segmentation is of excellent quality. The Moderate example shows a case with severe pathology which makes the upper part of the right lung (circled) appear disconnected. This region is not successfully segmented. The lobe segmentation images show fissure portions which are visible in the scan highlighted with white dotted lines. The Good lobe segmentation has a minor error in the segmentation of the right middle lobe which is not connected to the mediastinum. Segmentation is otherwise excellent. The Moderate example shows a case where the lobes are poorly segmented. The subject in this case suffered from emphysema making correct identification of the fissures very difficult. In the vessel segmentation row, the leftmost image received a score of Good since it is largely correct. Some regions of noise/high density parenchyma are erroneously segmented, however. This is most noticeable in the right dorsal lung. On the right, the Moderate example is a subject with visible pathology in the right lung. This pathology has a similar density to vessel and is therefore incorrectly segmented in this case. The registration example images show fissures high-lighted with dotted lines for clarity. The Good example has minor errors in fissure alignment, particularly in the right lung. Vessels are mainly well aligned with some minor exceptions. The Moderate example shows more severe misalignment of the fissures as well as the vessels. This was a case requiring a large deformation which has not been correctly resolved.

V.B. Correlation of quantitative measurements and pulmonary function test results

To provide an overview of whether the quantitative measures calculated from the thoracic CT data were in fact related to the GOLD stage, as defined by pulmonary function tests, the correlation coefficients between each quantitative measure and the known GOLD stage (0–4) of the subject were calculated. This verifies whether the measures calculated are meaningful in terms of airflow limitation or not.

Furthermore, the correlation between each of the quantitative measures and the component pulmonary function measures which are used to determine the GOLD stage (FEV1/FVC% and FEV1% predicted) was measured. This determines whether the ventilation measures calculated have greater correlation with one or other of these pulmonary function test scores, which may provide information of clinical interest.

TABLE III. Scores assigned by radiology experts to the 432 scans (216 scan pairs) for segmentation and registration results.

Number of scans	Excellent	Good	Moderate	Poor	Total
Lung segmentation	395	32	5	0	432
Lobe segmentation	252	120	54	6	432
Vessel segmentation	173	208	51	0	432
Registration (pairs)	59	129	26	2	216

V.C. Gold stage classification

A kNN classifier was used to attempt to classify subjects into the correct GOLD stage class based on feature sets extracted from the quantitative measures calculated. All experiments described are based on leave-one-out cross validation with k set empirically at 6. The classification was carried out on the subgroup of subjects where all segmentations and registrations were scored as satisfactory, as well as on the full group of subjects (disregarding the expert scores).

Nine sets of features were defined in order to determine whether particular feature types were more useful than others. Determining which features are more useful is done partly in order to learn how to tune the system to get the best results, but also to illustrate points of clinical interest, for example, the influence of certain lung regions compared to others, or the utility of the expiration scan compared with the inspiration scan. The feature sets used are listed below with feature labels corresponding to those listed in Table II.

1. Full set: All possible measures as described in Table II, each one calculated per lobe, per lung, and for the total lung volume (88 features, A1–A8…K1–K8)
2. No lobes: The per lung and total lung volume measures only. (33 features, A1–A3…K1–K3)
3. Only lobes: The per lobe measures only. (55 features, A4–A8…K4–K8)
4. Only upper lobes: The per lobe measures from the left upper and right upper lobes only. (22 features, A4–A5…K4–K5)
5. Only lower lobes: The lobe measures from the left lower and right lower lobes only. (22 features, A7–A8…K7–K8)
6. No registration: The measures extracted from the individual inspiration and expiration scans only, excluding those requiring registration. (24 features, A1–A8…C1–C8)
7. Only registration: The measures extracted from the registered data only. (64 features, D1–D8…K1–K8)
8. Insp scan only: The measures extracted from the inspiration scan only. (16 features, B1–B8…C1–C8)

TABLE IV. Numbers of scan pairs excluded by unsatisfactory (Moderate/Poor) scores in various categories. Pairs excluded in each category may be fewer than pairs with unsatisfactory scores since an unsatisfactory score in an earlier category would have excluded the pair at that stage.

Number of scan pairs	Moderate/Poor	Excluded	Remaining (/216)
Lung segmentation	5	5	211
Lobe segmentation	45	40	171
Vessel segmentation	43	28	143
Registration	28	17	126

TABLE V. The numbers of subjects at various GOLD stages in the database both before and after processing.

GOLD stage	0	1	2	3	4	Total
Before processing	49	49	50	42	26	216
Processed, scores satisfactory	24	29	35	22	16	126

9. Exp scan only: The measures extracted from the expiration scan only. (8 features, A1–A8)

VI. RESULTS

VI.A. Visual scoring

Scores provided by the radiology experts are shown in Table III for each of the segmentation and registration results. Scans pairs where both scans were scored Excellent or Good in all categories were retained as satisfactory while the remaining pairs were excluded. Table IV shows details of the pairs that were excluded based on unsatisfactory scores in various categories. Here, the numbers given are in terms of scan pairs rather than individual scans (i.e., exclusion of a scan pair may be due to a poor segmentation in only one scan from the pair). The number of pairs excluded may be less than the number of pairs which were unsatisfactory for a particular processing step, since if a scan pair was unsatisfactory in an earlier step it would have been excluded previously at that stage. The number of pairs remaining after each step is reported showing a final database of 126 scan pairs where all processing was scored as visually satisfactory. The division of these subjects into various GOLD stages is given in Table V. It can be seen that a reasonable balance of subjects in each GOLD stage is retained. The full database of 216 subjects will hereafter be referred to as the full group, while the group where all scores were either Excellent or Good (126 subjects) will be referred to as the satisfactory group.

VI.B. Correlation of quantitative measurements and pulmonary function test results

In Table VI, the correlation coefficients, r, between the quantitative measures calculated and the reference standard of the subject COPD GOLD stage are shown. Note that each measure is calculated per lobe, for the left lung, the right lung, and the overall lung volume, and correlation coefficients are given for each of these regions. All measures are based on the 126 subjects in the satisfactory group.

TABLE VI. Correlation coefficients between GOLD values and calculated quantitative measures for the 126 subjects in the satisfactory group. For convenience, the absolute values of correlation coefficients are displayed. Lobe names are abbreviated such that RUL implies "right upper lobe" and so on.

Measure	RUL	RML	RLL	LUL	LLL	Left	Right	Total
Perc850Exp	0.82	0.76	0.85	0.84	0.84	0.87	0.86	0.87
Perc910Insp	0.63	0.43	0.62	0.55	0.63	0.62	0.66	0.65
Perc950Insp	0.70	0.61	0.69	0.66	0.71	0.72	0.74	0.74
AvgSubtr	0.80	0.76	0.84	0.82	0.85	0.86	0.84	0.85
Avg Ratio	0.80	0.76	0.85	0.83	0.85	0.86	0.84	0.86
AvgVentil	0.65	0.54	0.77	0.70	0.74	0.75	0.75	0.76
AvgVentilBMC	0.56	0.49	0.69	0.60	0.67	0.66	0.65	0.66
MedSubtr	0.80	0.75	0.84	0.82	0.85	0.85	0.83	0.85
MedRatio	0.80	0.75	0.84	0.82	0.85	0.85	0.83	0.85
MedVentil	0.76	0.71	0.81	0.78	0.81	0.82	0.80	0.81
MedVentilBMC	0.71	0.65	0.77	0.73	0.77	0.77	0.74	0.76

TABLE VII. Correlation coefficients between measures calculated from CT and the pulmonary function measures of FEV1/FVC% and FEV1%predicted for the 126 subjects in the satisfactory group.

	FEV1/FVC%			FEV1%predicted		
Measure	Left	Right	Total	Left	Right	Total
Perc850Exp	0.91	0.89	0.91	0.86	0.86	0.86
Perc910Insp	0.69	0.71	0.71	0.53	0.58	0.56
Perc950Insp	0.79	0.79	0.80	0.69	0.72	0.71
AvgSubtr	0.86	0.84	0.85	0.89	0.88	0.89
AvgRatio	0.87	0.85	0.86	0.89	0.87	0.89
AvgVentil	0.74	0.73	0.74	0.83	0.82	0.83
AvgVentilBMC	0.65	0.64	0.65	0.72	0.72	0.73
MedSubtr	0.86	0.84	0.85	0.88	0.86	0.88
MedRatio	0.86	0.84	0.85	0.88	0.86	0.87
MedVentil	0.81	0.79	0.81	0.88	0.86	0.88
MedVentilBMC	0.76	0.74	0.75	0.83	0.81	0.82

The correlation coefficients between the ventilation measures and the pulmonary function measures of FEV1/FVC% and FEV1%predicted are shown in Table VII. For brevity, only the results for the left and right lung and the total lung volume are shown; however, similar patterns are seen in the per lobe measures.

VI.C. GOLD stage classification

The results of the experiments to classify subjects into the correct GOLD stages are shown in Table VIII. For each set of features that was used, we report (a) the percentage of

TABLE VIII. The results of the experiments to classify subjects into GOLD stages using a kNN classifier with k set empirically at 6 and using leave-one-out cross validation. Numbers in brackets imply the number of features used. Satisfactory group columns give the results for the 126 subjects where all segmentation and registration results were scored as satisfactory. Full group columns give the results for the full group of 216 subjects regardless of radiology scoring. "%correct" gives the percentage of subjects where COPD classification was correct. "%within 1 class" gives the percentage of subjects who were assigned either the correct class or a class neighboring the correct one.

	Satisfactory group (N = 126)		Full group (N = 216)	
Feature set (#)	% correct	%within 1 class	% correct	%within 1 class
Full feature set (88)	59	91	54	94
No lobes (33)	58	89	52	92
Only lobes (55)	60	90	57	93
Only upper lobes (22)	54	90	44	92
Only lower lobes (22)	67	96	56	93
No registration (24)	54	91	48	89
Only registration (64)	67	90	57	93
Insp scan only (16)	40	80	40	78
Exp scan only (8)	59	94	52	93

subjects correctly classified (% correct) and (b) the percentage of subjects that were either correctly classified or assigned a class neighboring the correct one (%within 1 class). Results are shown for the 126 subjects where all segmentation and registration scores were deemed satisfactory (satisfactory group) as well as for the full group of 216 subjects where segmentations and registrations were not necessarily scored as satisfactory (full group).

VII. DISCUSSION

VII.A. Workflow analysis

The first aim of this work was to make a detailed analysis of the prototype system for determining ventilation measurements from CT. Each segmentation and registration in the pipeline leading to the calculation of the ventilation measurements was examined and scored by a radiology expert. The results of this scoring are provided in Tables III and IV. This type of detailed analysis is essential in order to fully understand the system and the weak points which require improvement.

In the lung segmentation step the scores were extremely good with 395 of the 432 scans identified as having an Excellent lung segmentation. No cases were considered to be Poor, and only five scans were labeled Moderate. None of these five segmentations were identified as erroneous by the error-checking system in the hybrid lung segmentation algorithm. Figure 3 shows an example of a Moderate lung segmentation on the top right. In this case, the algorithm was unable to segment the lungs fully because of unusually severe pathology in the subject. In general, it is inevitable that any automatic segmentation or registration method will fail in a small number of the most difficult or unusual cases. As technology advances and automatic algorithms make their way into clinical practice, it is clear that any automated system should not only be successful in a high percentage of scans processed but also be able to identify and flag cases in which processing has been unsuccessful. Such cases might then be referred to a different system or to a clinician for manual intervention. The lung segmentation step in this system functions extremely well and requires only some additional focus on automatic error detection to adapt it for use in clinical practice.

Lobe segmentation was deemed satisfactory in 372 of the 432 scans processed with 252 of these being scored as Excellent. Only six cases were considered to have Poor lobe segmentation, while 54 were scored as Moderate. As previously noted, the lobe segmentation is dependent on the airway and fissure segmentations (in addition to lung segmentation). Segmenting any of these small structures is typically difficult, and especially so in this type of data, where the signal-to-noise ratio is frequently poor and in expiration data where the airways have collapsed and the parenchyma takes on a brighter appearance. The Moderate example of lobe segmentation shown in Fig. 3 illustrates a further difficulty in scans with severe emphysema or other pathology. Considering the quality and type of data processed in this work, the lobe segmentation results are very promising. However, as with all segmentation methods, we anticipate that there will always be a number of scans which cannot be successfully segmented and therefore suggest that future work will focus on automatic error detection in addition to enhancements to the segmentation algorithm.

Vessel segmentation in CT is typically a very difficult task, particularly for noisy data such as that included in the database used here. However, the results here are very encouraging with 381 of the 432 scans being scored as Excellent or Good. No vessel segmentations were considered to be Poor and just 51 were scored as Moderate. The vessel segmentation examples shown in Fig. 3 show some typical errors where pathological tissue or noise with bright intensities are segmented as vessel. The vessel segmentation used in this system could certainly be refined and improved, since at present it is partially based on a relatively crude thresholding method in order to avoid undersegmentation. This thresholding is responsible for the incorrect segmentation of bright structures in poor quality or pathological scans.

Registration results were scored as Excellent for 59 of the 216 pairs examined, and as Good for a further 129 pairs. Only two pairs were considered Poor, while 26 were Moderate. These results are remarkably good considering that accurate registration of breath-hold inspiration and expiration pairs remains a very difficult task due to the large nonuniform deformations encountered. Registration methods are constantly improving and results from some very promising algorithms have been published,[14,15] suggesting that future versions of this system may be able to improve upon the current registration success rate. However, it is also extremely important to consider developing a system where registration

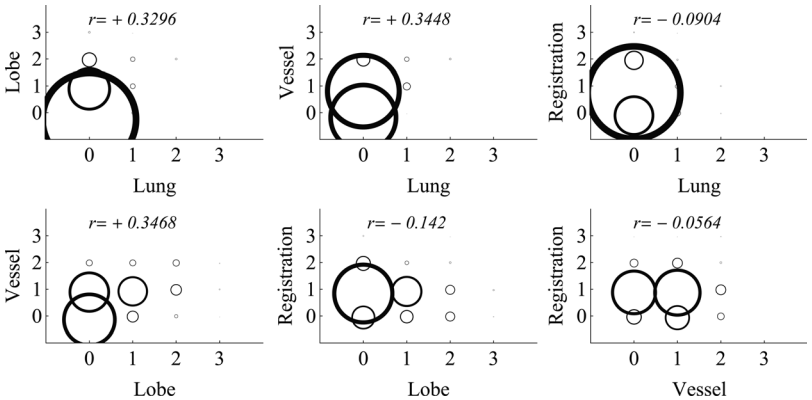

FIG. 4. An illustration of the correlation between the various segmentation and registration scores assigned by expert observers. Scores in each category are represented by 0 (excellent), 1 (good), 2 (moderate), and 3 (poor). The plot axis labels represent the defined tasks (segmentation of lung, lobe, and vessel and registration). The size of the circle drawn at a plot location indicates the number of scans which had those scores in those tasks. (For reference, the largest circle, shown on the first plot, represents 246 scans). The value "r" shown on each plot indicates the correlation of the scores between the two tasks.

failures can be automatically identified before the pipeline based on their results is allowed to continue processing.

Figure 4 illustrates the correlations between scores in the four tasks discussed above. It can be seen that there is no meaningful correlation between the scores in the various tasks. Although the other tasks depend on the lung segmentation, there are only five cases where the lung segmentation was not Good or Excellent, meaning that it rarely had an adverse affect on the subsequent processing. None of the plots show a large number of scans along the diagonal which, along with the poor correlation results shown, indicates that the scores are largely independent of each other.

This first part of our analysis indicates that individual steps in the pipeline perform well, particularly in consideration of the noisy and frequently pathological data being processed.

VII.B. Validation of quantitative measurements

As a further step toward analyzing the merit of the proposed system experiments were carried out to check the correlation of CT-derived measurements with those from pulmonary function tests. Furthermore, a kNN classifier was employed to assign GOLD stages to subjects based on the automatic ventilation measurements. The results from these experiments indicate whether or not the measures being calculated are meaningful for determining pulmonary function.

In Table VI, the correlation coefficients between the quantitative measures calculated and the GOLD stages of the subjects are shown. It is clear that all of the quantitative measures are correlated to some degree with the assigned GOLD stage, and many of them are highly correlated. In particular, the percentage of parenchyma below −850 HU in the expiration scan (Perc850Exp) shows excellent correlation in all regions of the lung, with an overall coefficient of 0.87 for the full lung volume. Similarly, the average and median subtraction values and ratio values (AvgSubtr, AvgRatio, MedSubtr, MedRatio) also have correlation coefficients between 0.85 and 0.86 for the combined left and right lungs.

Correlation coefficients, r, with the pulmonary function measures of FEV1/FVC% and FEV1%predicted are given in Table VII. Once again, very good correlation is seen in all experiments with a median r value of 0.84 over all values in the table. In fact, all correlation coefficients in both Tables VI and VII are found to be significant using t-tests with significance level $p < 0.001$ and a null hypothesis of no relationship between the variables. We conclude, therefore, that the CT measurements have a definite relationship with conventional pulmonary function testing.

The results of kNN classification experiments are shown in Table VIII. The ability to successfully classify COPD stages using the CT-derived features further reinforces the conclusion that these features are strongly related to pulmonary function testing. Classification results are extremely promising for such a complex multiclass problem where the reference standard is based on an entirely independent measurement. Using the 22 features derived from the lower lobes obtains the optimum results for the satisfactory group of 126 subjects. In this experiment, 67% of subjects are correctly classified while 96% are assigned either the correct class or a class neighboring it. Classification to a class neighboring the correct one is considered as a positive finding since the boundary between GOLD classes is based on a thresholding of measurements made on a continuous scale. An arbitrarily minor change in a subject's performance during spirometry may therefore result in a reclassification to a neighboring GOLD stage.

Comparing the classification results for the satisfactory group (N = 126) with those for the Full group (N = 216) whose segmentation and registration quality had not been taken into consideration, it can be seen that the percentage of correct classifications is not improved in any experiment by using the larger group. This is in spite of the fact that a larger sample size would be expected to provide improved results. The percentage of classifications correct to within 1 class is slightly improved in some cases when the larger group is used but disimproved in others. Overall, we conclude that performance is somewhat, although not considerably, enhanced by using only those scans where all segmentations and registrations are satisfactory.

VII.C. Most Effective quantitative measurements

The third aim of this work was to examine the quantitative measures calculated and analyze which of these appear to be most useful. Table VIII shows that measurements derived from registering the image pair (only registration) are more useful classification features (67% classified correctly) than measurements from the individual images only (no registration) (54% classified correctly). This suggests that registration based measurements add important value to the system. The correlation results in Table VI back up this theory, illustrating that the measurements derived from the subtraction and ratio images (AvgSubtr, AvgRatio, MedianSubtr, MedianRatio) all have excellent correlation with the GOLD stage values. It must also be noted, however, that the highest correlation value in Table VI is with the measurement Perc850Exp which is extracted from the expiration scan alone. GOLD stage classification using features from the expiration scan alone is also quite successful (59% correctly classified, see Table VIII), particularly considering there are only eight features used in this experiment. It may therefore be concluded that both registration-related features and features from the expiration scan are desirable to obtain optimal results. From Tables VI and VIII, the measurements from the inspiration scan alone appear to be the least useful of all those calculated, possibly since they do not show details such as air-trapping, which is an important characteristic of COPD. The correlation values for the ventilation measurements (Ventil and VentilBMC) are also relatively low.

In Table VII, correlation values with the pulmonary function measures of FEV1/FVC% and FEV1%predicted are shown. One interesting point to note from this table is that the measurements from the individual inspiration and expiration CT scans correlate best with the FEV1/FVC% data while the registration-related measurements have better correlation with the FEV1%predicted values. Although it cannot be conclusively stated based on this database alone, this suggests that the measurements from the individual scans (most

particularly the expiration scan) are useful in determining the presence or absence of COPD (FEV1/FVC% $\geq$ 70) while those from the registered scan pairs are more sensitive to the level of severity of the COPD (see Table I).

VII.D. Advantages of the proposed system

Having analyzed the system, identified its weak points and verified that it provides meaningful results, we further proposed to determine the potential advantages of this method over conventional spirometry and to highlight initial findings of clinical interest. One of the most important advantages of this system lies in its ability to make regional measurements of pulmonary function. Table VI shows that, in general, lobe based measurements correlate well with GOLD stage values, and therefore, it may be concluded that they offer meaningful information about pulmonary function at a regional level. Analyzing pulmonary function on a per lobe basis may become an extremely powerful tool in the future, improving knowledge about specific illnesses and appropriate treatment plans.

It can be seen in Table VIII that the measurements from the lower lobes (only lower lobes) are considerably better as classification features than those from the upper lobes (only upper lobes). In fact, the classification using only lower lobe information obtains better results than any other classification experiment listed in Table VIII. In addition Table VI shows a trend of higher correlation for the lower lobes than for the upper. These results strongly imply that the airflow in the lower lobes has a more substantial effect on the overall pulmonary function of the patient than that in the upper lobes. To our knowledge, this clinically important information has not been quantitatively proven in the past.

Although breath-hold inspiration is a much more frequently acquired scan than breath-hold expiration, we have shown that in terms of determining pulmonary function the expiration scan alone is more useful. The inspiration scan alone gives relatively poor performance in terms of correlation with GOLD stage (Table VI) and in the classification experiments (Table VIII). However, we have also shown that the combination of registered inspiration and expiration scans is extremely useful, and that for optimal performance, it is preferable to have both scans available (see Sec. VII C). This is important information for the future in terms of determining scanning protocols for assessing regional pulmonary function through CT.

VII.E. Limitations

As the gold standard in pulmonary function testing, spirometry is the best reference standard against which to compare the system developed in this work. However, we recognize that spirometric measures have some limitations, in that their accuracy and reproducibility relies upon the ability of the subject to follow instructions carefully as well as his health and condition at the moment of testing. The COPD classification system is also slightly flawed in that classes are assigned based on the thresholding of measurements made on a continuous scale. An arbitrarily small change in the subject's pulmonary performance may therefore result in a different COPD classification.

The quality of much of the CT data used in this work may also be considered relatively poor as it is acquired using an ultralow dose protocol. For the 216 scan pairs considered, the beam current (c) information was analyzed where available (209 scan pairs). Only 19 of these scan pairs had an average beam current above 30 mAs; 172 pairs had an average c of 25 mAs. Those scan pairs where a higher beam current was used are drawn from clinical practice, while the majority of the low-dose data is drawn from the Nelson trial (of asymptomatic subjects), meaning that for our data set a high value of c is likely to imply a more severe stage of COPD. For example, 25% of the 16 subjects with average c of 75 mAs have a GOLD class of 4, which is true of only 9% of the 172 subjects with average c of 25 mAs. The data used here are therefore insufficient to provide a full analysis of how the scan dose affects the outcome of our system. However, we believe it is likely that using data with a better signal-to-noise ratio would improve the system results. Nonetheless, it is advantageous to have shown that the system is also capable of working on low-dose data which can be acquired with less risk to the subject.

VIII. CONCLUSIONS

This work demonstrates a prototype system for assessing ventilation in a regional, quantitative manner based on the appearance of the lungs in CT at inspiration and expiration. Experiments are carried out on a database of 216 subjects with well distributed varying levels of COPD and compared to the conventional reference standard of spirometry testing. The method does not require any contrast agent and is tested on low and ultralow dose CT data, verifying that it can be implemented with minimal radiation exposure to the patient.

A thorough analysis of the pipeline leading to the calculation of ventilation measurements shows that individual steps are very successful. Segmentation of lungs, vessels, and lobes was satisfactory in 99%, 88%, and 86% of cases, respectively, while registration succeeded in 87% of subjects.

Eleven different ventilation measurements, extracted from the individual CT scans and from the registered scans, are calculated on 8 defined lung regions from individual lobes up to total lung volume. All measurements on all regions are demonstrated to correlate with spirometry and GOLD stage information. The mean correlation coefficient, r, for the ventilation measures calculated over the total lung volume correlated with the subject GOLD stage was 0.79. Using a selection of the CT-derived features in a kNN classifier to assign subject GOLD stages resulted in successful classification in 67% of subjects with a further 29% of subjects assigned a class neighboring the correct one.

The most successful measurements for determining pulmonary function are obtained by thresholding (-850 HU) on the expiration scan, and by direct comparison (subtraction/ratio) of the registered scan pair values. The inspiration scan alone gives a relatively weak performance in assessing airflow.

With further development and testing on additional databases including different pulmonary pathologies, there is enormous potential for findings which may be key to a better understanding of the mechanisms of pulmonary function and disease.

ACKNOWLEDGMENTS

The authors would like to thank Marius Staring, Leiden University Medical Center, for assistance with configuration of the *elastix* software.

APPENDIX: SCORING GUIDELINES FOR RADIOLOGY EXPERTS

The radiology experts were asked to score lung, lobe, and vessel segmentations as well as registrations of inspiration and expiration scans. Proprietary software was created for this purpose.

1. Segmentation scoring

To score segmentations, the expert could scroll through the entire scan with the segmentation overlaid in semitranslucent color. This overlay could be removed and replaced at any time if the expert wished to see some scan details more clearly. The guidelines provided for scoring were as follows:

- Excellent: 95% of the segmentation is correct
- Good: 80%–95% of the segmentation is correct
- Moderate: 50%–80% of the segmentation is correct
- Poor: 50% or less of the segmentation is correct

2. Registration scoring

For registration scoring, the expert scrolled through the original scan and the registered scan simultaneously. He was requested to first scroll through the slices and get a general idea of whether alignment looked good or not. Next, he was asked to select some example slices and examine them in more detail to see whether fissures and vessels were well aligned. A crosshairs tool was provided which when hovered over a point in one scan showed the corresponding point in the other. This allowed the expert to see whether specific small structures were aligned. The expert was instructed to continue selecting example slices from various regions of the image until satisfied that he had a good overview of the registration quality for both fissures and vessels. Guidelines for scoring were provided as follows:

- Excellent: If almost every structure (vessel or fissure, depending on the category) checked was very well aligned
- Good: If the structures were well aligned in most places, but there were a few locations with misalignments of up to a few mm
- Moderate: If there were severe misalignments, and if these were present in multiple locations throughout the image
- Poor: If the images were completely misaligned

[a]Author to whom correspondence should be addressed. Electronic mail: keelinm@gmail.com

[1]D. S. Gierada, T. K. Pilgram, B. R. Whiting, C. Hong, A. J. Bierhals, J. H. Kim, and K. T. Bae, "Comparison of standard- and low-radiation-dose CT for quantification of emphysema," Am. J. Roentgenol. **188**(1), 42–47 (2007).
[2]T. Kubo, P. J. Paul Lin, W. Stiller, M. Takahashi, H. U. Kauczor, Y. Ohno, and H. Hatabu, "Radiation dose reduction in chest CT: A review," Am. J. Roentgenol. **190**(2), 335–343 (2008).
[3]M. Takahashi, W. M. Maguire, M. Ashtari, A. Khan, Z. Papp, R. Alberico, W. Campbell, T. Eacobacci, and P. G. Herman, "Low-dose spiral computed tomography of the thorax: Comparison with the standard-dose technique," Invest. Radiol. **33**(2), 68–73 (1998).
[4]X. Zhu, J. Yu, and Z. Huang, "Low-dose chest CT: Optimizing radiation protection for patients," Am. J. Roentgenol. **183**(3), 809–816 (2004).
[5]H. Arakawa, H. Niimi, Y. Kurihara, Y. Nakajima, and W. R. Webb, "Expiratory high-resolution CT: Diagnostic value in diffuse lung diseases," Am. J. Roentgenol. **175**(6), 1537–1543 (2000).
[6]H. Arakawa and W. R. Webb, "Expiratory high-resolution CT scan," Radiol. Clin. North Am. **36**(1), 189–209 (1998).
[7]M. Nishino, G. R. Washko, and H. Hatabu, "Volumetric expiratory HRCT of the lung: Clinical applications," Radiol. Clin. North Am. **48**(1), 177–183 (2010).
[8]E. J. R. van Beek and E. A. Hoffman, "Functional imaging: CT and MRI," Clin. Chest Med. **29**(1), 195–216 (2008).
[9]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).
[10]E. J. R. van Beek, G. McLennan, and E. A. Hoffman, "Ventilation imaging using computed tomography," *Imaging Decisions MRI* (Wiley, 2004), Vol. 8, p. 15–23.
[11]H. U. Kauczor, J. Hast, C. P. Heussel, J. Schlegel, P. Mildenberger, and M. Thelen, "CT attenuation of paired HRCT scans obtained at full inspiratory/expiratory position: Comparison with pulmonary function tests," Eur. Radiol. **12**(11), 2757–2763 (2002).
[12]H. U. Kauczor, C. P. Heussel, B. Fischer, R. Klamm, P. Mildenberger, and M. Thelen, "Assessment of lung volumes using helical CT at inspiration and expiration: Comparison with pulmonary function tests," Am. J. Roentgenol. **171**(4), 1091–1095 (1998).
[13]J. Zaporozhan, S. Ley, R. Eberhardt, O. Weinheimer, S. Iliyushenko, F. Herth, and H. U. Kauczor, "Paired inspiratory/expiratory volumetric thin-slice CT scan for emphysema analysis: Comparison of different quantitative evaluations and pulmonary function test," Chest **128**(5), 3212–3220 (2005).
[14]K. Murphy, B. van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. E. Christensen, V. Garcia, T. Vercauteren, N. Ayache, O. Commowick, G. Malandain, B. Glocker, N. Paragios, N. Navab, V. Gorbunova, J. Sporring, M. de Bruijne, X. Han, M. P. Heinrich, J. A. Schnabel, M. Jenkinson, C. Lorenz, M. Modat, J. R. McClelland, S. Ourselin, S. E. A. Muenzing, D. De Nigris, D. L. Collins, T. Arbel, M. Peroni, R. Li, G. C. Sharp, A. Schmidt-Richberg, J. Ehrhardt, R. Werner, D. Smeets, D. Loeckx, G. Song, N. Tustison, B. Avants, J. C. Gee, M. Staring, S. Klein, B. C. Stoel, M. Urschler, M. Werlberger, J. Vandemeulebroucke, S. Rit, D. Sarrut, and J. P. W. Pluim, "Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge," IEEE Trans. Med. Imaging **30**(11), 1901–2000 (2011).
[15]K Murphy, B. van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X Deng, and J. P. W. Pluim, "Evaluation of methods for pulmonary image registration: The EMPIRE10 study," *Medical Image Analysis for the Clinic: A Grand Challenge* (2010), pp. 11–22.
[16]J. M. Reinhardt, K. Ding, K. Cao, G. E. Christensen, E. A. Hoffman, and S. V. Bodas, "Registration-based estimates of local lung tissue expansion compared to xenon CT measures of specific ventilation," Med. Image Anal. **12**(6), 752–763 (2008).
[17]T. Guerrero, K. Sanders, E. Castillo, Y. Zhang, L. Bidaut, T. Pan, and R. Komaki, "Dynamic ventilation imaging from four-dimensional computed tomography," Phys. Med. Biol. **51**(4), 777–791 (2006).
[18]K. Ding, K. Cao, R. E. Amelon, G. E. Christensen, M. L. Raghavan, and J. M. Reinhardt, "Comparison of intensity- and Jacobian-based estimates of lung regional ventilation," *Third International Workshop on Pulmonary Image Analysis* (CreateSpace, Beijing, 2010), pp. 49–60.
[19]R. Castillo, E. Castillo, J. Martinez, and T. Guerrero, "Ventilation from four-dimensional computed tomography: Density versus Jacobian methods," Phys. Med. Biol. **55**(16) (2010).
[20]C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," PLoS Med. **3**(11), e442 (2006).
[21]K. F. Rabe, S. Hurd, A. Anzueto, P. J. Barnes, S. A. Buist, P. Calverley, Y. Fukuchi, C. Jenkins, R. Rodriguez-Roisin, C. van Weel, and J. Zielinski, "Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary," Am. J. Respir. Crit. Care Med. **176**(6), 532–555 (2007).
[22]R. Pellegrino, G. Viegi, V. Brusasco, R. O. Crapo, F. Burgos, R. Casaburi, A. Coates, C. P M van der Grinten, P. Gustafsson, J. Hankinson, R. Jensen, D. C. Johnson, N. MacIntyre, R. McKay, M. R. Miller, D. Navajas, O. F. Pedersen, and J. Wanger, "Interpretative strategies for lung function tests," Eur. Respir. J. **26**(5), 948–968 (2005).
[23]D. M. Xu, H. Gietema, H. de Koning, R. Vernhout, K. Nackaerts, M. Prokop, C. Weenink, J. Lammers, H. Groen, M. Oudkerk, and

R. van Klaveren, "Nodule management protocol of the NELSON randomised lung cancer screening trial," Lung Cancer **54**(2), 177–184 (2006).
[24]E. M. van Rikxoort, B. J. de Hoop, M. A. Viergever, M. Prokop, and B. van Ginneken, "Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection," Med. Phys. **36**(7), 2934–2947 (2009).
[25]E. M. van Rikxoort, B. van Ginneken, M. Klik, and M. Prokop, "Supervised enhancement filters: Application to fissure detection in chest CT scans," IEEE Trans. Med. Imaging **27**(1), 1–10 (2008).
[26]B. van Ginneken, W. Baggerman, and E. M. van Rikxoort, "Robust segmentation and anatomical labeling of the airway tree from thoracic CT scans," *Medical Image Computing and Computer Assisted Intervention* (LNCS, New York, 2008), Vol. 5241, pp. 219–226.
[27]E. M. van Rikxoort, M. Prokop, B. J. de Hoop, M. A. Viergever, J. P. W. Pluim, and B. van Ginneken, "Automatic segmentation of pulmonary lobes robust against incomplete fissures," IEEE Trans. Med. Imaging **29**(6), 1286–1296 (2010).
[28]P. Lo, J. Sporring, H. Ashraf, J. J. H. Pedersen, and M. de Bruijne, "Vessel-guided airway tree segmentation: A voxel classification approach," Med. Image Anal. **14**(4), 527–538 (2010).
[29]A. M. R. Schilham, B. van Ginneken, H. Gietema, and M. Prokop, "Local noise weighted filtering for emphysema scoring of low-dose CT images," IEEE Trans. Med. Imaging **25**, 451–463 (2006).
[30]S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "Elastix: A toolbox for intensity-based medical image registration," IEEE Trans. Med. Imaging **29**(1), 196–205 (2010).
[31]http:// elastix.isi.uu.nl
[32]M. Staring, S. Klein, J. H. C. Reiber, W. J. Niessen, and B. C. Stoel, "Pulmonary image registration with elastix using a standard intensity-based algorithm," *Medical Image Analysis for the Clinic: A Grand Challenge* (CreateSpace, Beijing, 2010), pp. 73–79.
[33]T. Yamamoto, S. Kabus, T. Klinder, C. Lorenz, J. von Berg, T. Blaffert, B. W. Loo, Jr., and P. J. Keall, "Investigation of four-dimensional computed tomography-based pulmonary ventilation imaging in patients with emphysematous lung regions," Phys. Med. Biol. **56**(7) (2011).
[34]T. Yamamoto, S. Kabus, T. Klinder, J. von Berg, C. Lorenz, B. W. Loo, and P. J. Keall, "Four-dimensional computed tomography pulmonary ventilation images vary with deformable image registration algorithms and metrics," Med. Phys. **38**, 1348–1358 (2011).
[35]T. Yamamoto, S. Kabus, J. von Berg, C. Lorenz, and P. J. Keall, "Impact of four-dimensional computed tomography pulmonary ventilation imaging-based functional avoidance for lung cancer radiotherapy," Int. J. Radiat. Oncol., Biol., Phys. **79**, 279–288 (2011).
[36]A. Busacker, J. D. Newell, T. Keefe, E. A. Hoffman, J. Cook Granroth, M. Castro, S. Fain, and S. Wenzel, "A multivariate analysis of risk factors for the air-trapping asthmatic phenotype as measured by quantitative CT analysis," Chest **135**(1), 48–56 (2009).
[37]D. A. Lynch and J. D. Newell, "Quantitative imaging of COPD," J. Thorac. Imaging **24**(3), 189–194 (2009).
[38]B. A. Simon, "Regional ventilation and lung mechanics using x-ray CT," Acad. Radiol. **12**, 1414–1422 (2005).
[39]R. Brower, R. A. Wise, C. Hassapoyannes, B. Bromberger-Barnea, and S. Permutt, "Effect of lung inflation on lung blood volume and pulmonary venous flow," J. Appl. Physiol. **58**(3), 954–963 (1985).