To appear in IEEE Transactions on Medical Imaging; http://dx.doi.org/10.1109/TMI.2012.2209674

Extraction of Airways from CT (EXACT'09)

Pechin Lo, Bram van Ginneken, Joseph M. Reinhardt, Tarunashree Yavarna, Pim A. de Jong, Benjamin Irving, Catalin Fetita, Margarete Ortner, Rômulo Pinho, Jan Sijbers, Marco Feuerstein, Anna Fabijańska, Christian Bauer, Reinhard Beichel, Carlos S. Mendoza, Sheikh Zayed, Rafael Wiemker, Jaesung Lee, Anthony P. Reeves, Silvia Born, Oliver Weinheimer, Eva M. van Rikxoort, Juerg Tschirren, Ken Mori, Benjamin Odry, David P. Naidich, Ieneke Hartmann, Eric A. Hoffman, Mathias Prokop, Jesper H. Pedersen, Marleen de Bruijne

P. Lo is with the Image Group, Department of Computer Science, University of Copenhagen, Denmark and with the UCLA Thoracic Imaging Research Group, Department of Radiology, University of California, Los Angeles, USA

B. van Ginneken is with the Diagnostic Image Analysis Group, Radboud University Nijmegen Medical Centre, The Netherlands and the Image Sciences Institute, University Medical Center Utrecht, The Netherlands

J. M. Reinhardt is with the Department of Biomedical Engineering, The University of Iowa, USA

T. Yavarna is with the Dept. of Biomedical Engineering, The University of Iowa, USA

P. A. de Jong is with University Medical Center Utrecht, Utrecht, The Netherlands

B. Irving is with University College London, WC1E 6BT, UK

C. Fetita is with Institut TELECOM / Telecom SudParis, Evry, and with MAP5 CNRS UMR 8145, France

M. Ortner is with Institut TELECOM / Telecom SudParis, Evry, and with MAP5 CNRS UMR 8145, France

R. Pinho is with the Univeristy of Lyon, Léon Bérard Cancer Centre, Lyon, France

J. Sijbers is with the University of Antwerp, Belgium

M. Feuerstein is with microDimensions, Technische Universität München, Germany

A. Fabijańska is with the Department of Computer Engineering, Technical University of Lodz, Poland

C. Bauer is with the Institute for Computer Graphics and Vision, Graz University of Technology, Austria and the Department of Electrical and Computer Engineering, Dept. of Electrical and Computer Engineering, USA

R. Beichel is with the Department of Electrical and Computer Engineering, The University of Iowa and the Department of Internal Medicine, The University of Iowa

C. S. Mendoza is with Department of Signal Processing and Communications, Universidad de Sevilla, Spain

S. Zayed is with Institute for Pediatric Surgical Innovation Children's National Medical Center, Washington DC, USA

R. Wiemker is with Philips Research Laboratories Hamburg, Germany

J. Lee is with the School of Electrical and Computer Engineering, Cornell University, USA

A. P. Reeves is with the School of Electrical and Computer Engineering, Cornell University, USA

S. Born is with Visual Computing, ICCAS, Universität Leipzig, Germany

O. Weinheimer is with the Department of Diagnostic and Interventional Radiology, Johannes Gutenberg University of Mainz, Germany

E. M. van Rikxoort is with the Diagnostic Image Analysis Group, Radboud University Nijmegen Medical Centre, The Netherlands

J. Tschirren is with VIDA Diagnostics, Inc., USA

K. Mori is with Information and Communications Headquarters, Nagoya University, Hirotsugu Takabatake, Minami Sanjo Hospital, Japan

B. Odry is with Corporate Research, Siemens Corporation, USA

D. P. Naidich is with the Department of Radiology, New York University Medical Center, USA

I. Hartmann is with Department of Radiology, Erasmus MC - University Medical Center Rotterdam, The Netherlands

E. A. Hoffman is with the Department of Radiology, The University of Iowa, USA

M. Prokop is with Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands and University Medical Center Utrecht, Utrecht, The Netherlands

J. H. Pedersen is with the Department of Cardio Thoracic Surgery, Rigshospitalet - Copenhagen University Hospital, Denmark

M. de Bruijne is with the Image Group, Department of Computer Science, University of Copenhagen, Denmark and Biomedical Imaging Group Rotterdam, Departments of Radiology & Medical Informatics, Erasmus MC - University Medical Center Rotterdam, The Netherlands

***Abstract*—This paper describes a framework for establishing a reference airway tree segmentation, which was used to quantitatively evaluate fifteen different airway tree extraction algorithms in a standardized manner. Because of the sheer difficulty involved in manually constructing a complete reference standard from scratch, we propose to construct the reference using results from all algorithms that are to be evaluated. We start by subdividing each segmented airway tree into its individual branch segments. Each branch segment is then visually scored by trained observers to determine whether or not it is a correctly segmented part of the airway tree. Finally, the reference airway trees are constructed by taking the union of all correctly extracted branch segments. Fifteen airway tree extraction algorithms from different research groups are evaluated on a diverse set of twenty chest computed tomography (CT) scans of subjects ranging from healthy volunteers to patients with severe pathologies, scanned at different sites, with different CT scanner brands, models, and scanning protocols. Three performance measures covering different aspects of segmentation quality were computed for all participating algorithms. Results from the evaluation showed that no single algorithm could extract more than an average of 74% of the total length of all branches in the reference standard, indicating substantial differences between the algorithms. A fusion scheme that obtained superior results is presented, demonstrating that there is complementary information provided by the different algorithms and there is still room for further improvements in airway segmentation algorithms.**

***Index Terms*—Pulmonary airways, computed tomography, segmentation, evaluation.**

I. Introduction

THE segmentation of airway trees in chest volumetric computed tomography (CT) scans plays an important role in the analysis of lung diseases. One application of airway tree segmentation is in the measurement of airway lumen and wall dimensions, which have been shown to correlate well with the presence of chronic obstructive pulmonary disease (COPD) [1], [2]. As the lungs are subdivided anatomically based on the airway tree, airway tree segmentation is also a useful input for other segmentation tasks such as segmentation of lobes [3], [4] and pulmonary segments [5], [6]. Airway segmentation is also a prerequisite for virtual bronchoscopy, which has increasingly been used to facilitate planning and guidance of bronchoscopic interventions [7], [8].

Several automated methods have been proposed to segment the airway tree from CT images. Evaluation of these methods has been problematic. Manual segmentation of airways is a diff cult and very time consuming task due to the complexity of the 3D structure of the airway tree. In addition,

low contrast in the peripheral branches may make manual detection, inevitably performed in 2D views, unreliable. Most methods have been evaluated qualitatively based on visual inspection or were compared quantitatively to more basic techniques such as region growing [9]–[16]. Some authors performed manual evaluation without constructing a ground truth segmentation. Tschirren et al. assessed the detection rate of their algorithm by manually assigning anatomical labels to detected branches [11], while Fetita et al. compared the number of automatically detected branches to the number of bronchial sections detected manually [17]. Other authors compared their results to segmentations obtained interactively, e.g., by region growing with manually selected thresholds [13], [15], or by manually removing "leaks" from the results of their proposed methods [18]–[20]. Graham et al. [8] obtained a ground truth for three airway trees from thin slice CT scans (1112 branches in total) using an interactive live-wire segmentation method for evaluation purposes. A drawback of such interactively obtained segmentations is that they may be biased to the algorithms used in their construction and are thus less suitable for comparing different methods. Although not very common, in some cases, a ground truth was constructed fully manually for evaluation. In [21]–[23], a single reference image is manually segmented, while Aykac et al. manually segment eight scans (471 branches in total) with 3 mm slices [24]. Because of the time required for manual annotation in these studies, evaluation was restricted to a small number of cases and inter-observer agreement was not studied.

The aim of this paper is to develop a framework to establish a reference airway tree segmentation that can be used to evaluate different airway tree extraction algorithms in a standardized manner. We believe that such standardized comparison of different algorithms is critical for future development, as the weaknesses of the different algorithms can be identified and possibly improved upon. Because of the sheer difficulty in manually establishing a complete reference standard from scratch, we propose to construct the reference from the results of the algorithms being evaluated. Segmented airway trees are first subdivided into their individual branches. These individual branches are then visually scored by trained observers and correctly segmented branches are retained, while incorrectly segmented branches are rejected. Airway segmentations produced by different algorithms on the same image will overlap to a certain extent. Therefore, the branch inspection process can be accelerated by automatically accepting branches that overlap with previously accepted branches. Finally, the reference standard is computed as the union of all accepted branches.

We use forty scans from eight different institutions. Scans were obtained under various acquisition conditions and with different scanners, at full inspiration or full expiration, and with a variety of pathological abnormalities. The first twenty scans are designated as a training set, and can be used to train and optimize algorithms. The remaining twenty scans are used as a testing set to evaluate the different algorithms.

The evaluation is designed to only take into consideration the depth of the airway trees extracted by an algorithm. We do not take the exact airway shape and dimensions into account: a branch is said to be correct as long as there is no significant leakage outside the airway walls.

This paper is based on the results of a comparative study that was organized at the 2nd International Workshop on Pulmonary Image Analysis[1], which was held in conjunction with the 12th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2009). Invitations were sent out to several mailing lists and to authors of published papers on airway tree segmentation. A total of 22 teams registered to download the data, and 15 teams [14], [25]–[38] submitted their results. This paper is based on the results of these 15 algorithms and as such presents a thorough, though not exhaustive, comparison of currently available algorithms. Ten teams [14], [25], [26], [28], [30], [32], [33], [35]–[37] submitted to the fully automated category and five teams [27], [29], [31], [34], [38] submitted to the semi-automated category. All results were used to establish the reference standard.

The evaluation results of the fifteen algorithms are the same as those reported in [39] and on the EXACT'09 website[2]. In this work, we thoroughly investigate algorithm performance by estimating the number of branches missed in our reference standard and by including local sensitivity analysis up to the segmental level, and we study the improvements that can be obtained by combining the results produced by different algorithms in a fusion framework.

II. Data

A total of 75 chest CT scans were contributed by eight different institutions. The scans were acquired with several different CT scanner brands and models, using a variety of scanning protocols and reconstruction parameters. The conditions of the scanned subjects varied widely, ranging from healthy volunteers to patients showing severe abnormalities in the airways or lung parenchyma. From the contributed scans, we selected forty scans for this study; a training set and a testing set of twenty scans each. All files were completely anonymized. An equal number of scans of similar quality, acquired at the same institutions and with similar protocols were included in both the training and testing sets, with no scans of the same subject included in both sets. We did not ensure that scans with similar anomalies were included in the training and testing sets. However, as the scans from both sets were from the same trial or clinical studies, it is likely that the scans from both sets were similar in terms of anomalies as well.

The images in the training set were named CASE01 through CASE20, and the images in the testing set were named CASE21 through CASE40. Table I presents acquisition parameters, a visual scoring of noise level, and a brief report of anomalies provided by a chest radiologist for the twenty test cases.

III. Airway branch scoring

This section describes how each airway branch segment is evaluated. We first describe how an airway tree segmentation

[1]See http://www.lungworkshop.org/2009/

[2]See http://image.diku.dk/exact/

To appear in IEEE Transactions on Medical Imaging; http://dx.doi.org/10.1109/TMI.2012.2209674

TABLE I

ACQUISITION PARAMETERS OF THE 20 TEST CASES. SLICE THICKNESS (T) IS GIVEN IN MM. TUBE VOLTAGE (TV) IS GIVEN IN KVP. AVERAGE TUBE CURRENT (TC) IS GIVEN IN MA. THE LEVEL OF INSPIRATION (LI) INDICATES WHETHER THE SCAN IS ACQUIRED AT FULL INSPIRATION (I) OR FULL EXPIRATION (E) WITH BREATH-HOLD. CONTRAST (C) INDICATES WHETHER INTRAVENOUS CONTRAST WAS USED DURING ACQUISITION ("Y" FOR YES AND "N" FOR NO). PERCEIVED RECONSTRUCTION (R) INDICATES WHETHER THE SCAN WAS RECONSTRUCTED USING A SOFT (S), MIDDLE (M) OR HARD (H) RECONSTRUCTION KERNEL, BASED ON VISUAL INSPECTION. THE NOISE LEVEL (N) OF THE SCAN IS SCORED BY VISUAL INSPECTION AS HIGH (H), MIDDLE (M) OR LOW (L). * INDICATES THAT A SCAN IS FROM THE SAME SUBJECT AS THE PRECEDING SCAN.

	T	Scanner	Kernel	TV	TC	LI	C	R	N	Anomalies
CASE21	0.6	Siemens Sensation 64	B50f	120	200.0	E	N	H	H	None
CASE22*	0.6	Siemens Sensation 64	B50f	120	200.0	I	N	H	H	None
CASE23	0.75	Siemens Sensation 64	B50f	120	200.0	I	N	H	M	None
CASE24	1	Toshiba Aquilion	FC12	120	10.0	I	N	M	H	Small lung nodule
CASE25*	1	Toshiba Aquilion	FC10	120	150.0	I	N	M	M	Small lung nodule
CASE26	1	Toshiba Aquilion	FC12	120	10.0	I	N	M	H	Intrafissural fluid
CASE27*	1	Toshiba Aquilion	FC10	120	150.0	I	N	M	M	Lympheadenopathy, bronchial wall thickening, airway collapse, septal thickening, intrafissural fluid
CASE28	1.25	Siemens Volume Zoom	B30f	120	348.0	I	Y	M	L	None
CASE29*	1.25	Siemens Volume Zoom	B50f	120	348.0	I	Y	M	L	None
CASE30	1	Philips Mx8000 IDT 16	D	140	120.0	I	N	M	M	Diffuse ground glass
CASE31	1	Philips Mx8000 IDT 16	D	140	120.0	I	N	M	L	Diffuse emphysema
CASE32	1	Philips Mx8000 IDT 16	D	140	120.0	I	N	M	L	Pleural plaques, mucus plug right lower lobe, few nodules
CASE33	1	Siemens Sensation 16	B60f	120	103.6	I	N	H	H	Mild bronchiectasis, mucus plugging, tree-in-bud pattern/small infiltrates
CASE34	1	Siemens Sensation 16	B60f	120	321.0	I	N	H	M	Mild bronchiectasis, mucus plugging, tree-in-bud pattern/small infiltrates
CASE35	0.625	GE LightSpeed 16	Standard	120	411.5	I	N	M	M	None
CASE36	1	Philips Brilliance 16P	C	120	206.0	I	N	S	L	Bronchiectasis, bronchial wall thickening, mucus plugs, infiltrates
CASE37	1	Philips Brilliance 16P	B	140	64.0	I	N	M	M	None
CASE38*	1	Philips Brilliance 16P	C	120	51.0	E	N	M	H	Air trapping
CASE39	1	Siemens Sensation 16	B70f	100	336.7	I	Y	H	H	Extensive bronchiectasis, many infiltrates and atelectasis, tree-in-bud, mucus plugging, central airway distortion
CASE40	1	Siemens Sensation 16	B70s	120	90.6	I	N	H	L	Extensive areas with ground glass

is subdivided into its individual branch segments. Next, we explain how a branch segment is presented to a human observer for visual assessment. The different labels used for scoring are then detailed. The rules to determine whether a branch segment requires visual assessment or can be accepted automatically are then introduced. Finally, we end this subsection by describing how the human observers were trained.

A. Subdividing an airway tree into branches

To enable evaluation of individual branches, an airway tree is first subdivided into its branches by a wave front propagation algorithm that detects bifurcations, as described in [40]. The key concept is that a wave front, propagating through a tree structure, remains connected until it encounters a bifurcation, and side branches can thus be detected as disconnected components in the wave front.

The front is propagated using the fast marching algorithm [41], [42], with a speed function that is equal to one inside and zero outside the segmented structure, thus limiting the front to only propagate within the segmented structure. The number of disconnected components is monitored by applying connected component analysis to the "trial" points in the front each time the front has moved a distance equal to the average distance between two voxels. If the front contains multiple disconnected components, the propagation proceeds by starting from the individual detected components and growing into the child branches. The process ends after the complete segmentation has been evaluated. During the front propagation, the centroid of the front is stored at every step to obtain the centerlines. Figure 1 illustrates the steps involved in the airway tree subdivision.

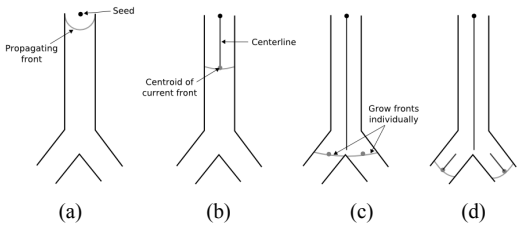


Fig. 1. Illustration of how an airway tree is subdivided into individual branches. (a) A seed point is placed at the root of a tree to initiate a front propagation process. (b) The centroid of the propagating front is stored as centerline during propagation. (c) The propagation is stopped when the front splits at a bifurcation, and new seeds are obtained from the individual split fronts. (d) The front propagation process is repeated from the new seed points.

B. Display

Visual assessment of each branch is conducted by displaying a fixed number of slices through the branch at different positions and orientations. Two different views are used to obtain the slices: a reformatted view that straightens the centerline of a branch segment, and a reoriented view that rotates the branch segment such that its main axis coincides with the x-axis.

Eight slices are extracted from the reformatted view. A schematic view of the slices are shown in Figure 2(b). The first four slices (A1, A2, A3 and A4) are taken perpendicular to the centerline, distributed evenly from the start to the end of

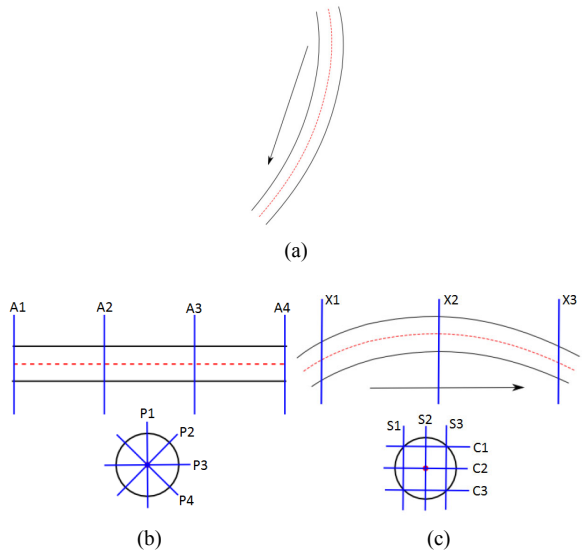


Fig. 2. Schematics showing the (a) original airway, (b) reformatted and (c) reoriented views. The arrow is the main orientation of the airway and the cut planes are shown in blue.

the centerline. The remaining four slices (P1, P2, P3, and P4) are taken along the centerline, at cut planes that are angled at 0°, 45°, 90° and 135°.

For the reoriented view, nine slices are extracted, consisting of three slices from each of the axial, sagittal, and coronal planes. Figure 2(c) presents a schematic view of the slices. For the slices in the sagittal (S1, S2 and S3) and coronal (C1, C2, and C3) planes, the slices are placed at 15%, 50%, and 85% of the branch width measured along the axis normal to the plane. On the axial plane (X1, X2 and X3), the slices are placed at 5%, 50%, and 95% of the branch length.

The segmentation is shown as a colored overlay on these slices. The user can toggle between the different views and toggle the overlay on and off for better assessment of the underlying structure. Figure 3 shows examples of the two views for a correctly segmented branch and a branch where the segmentation has leaked outside of the branch. The various slice display parameters for the two views were determined based on a trial study, in which we found the best trade off between the accuracy of the human observer's scores and the time required to score a single airway branch segment.

C. Scoring of branches by trained observers

The process of scoring the individual branches of all submitted segmentations was distributed among ten trained observers through a web-based system. The observers were all medical students who were familiar with CT and chest anatomy.

Using the slice display described in Section III-B, observers were asked to assign to each branch one of the following four labels: "correct", "partly wrong", "wrong" or "unknown". A branch is scored as "correct" if it does not have leakage outside the airway wall. "Partly wrong" is assigned to a branch if part of the branch lies well within the airway lumen and the remaining part of it lies outside the airway wall. A branch is "wrong" if it does not contain airway lumen at all. The "unknown" label is used when the observers are unable to determine whether a branch is an airway or not.

The scoring of each branch is performed in two phases. In phase one, two observers are assigned to score a branch. If both observers assigned the same label, the scoring is complete. Otherwise, the scoring proceeds to phase two, where three new observers are assigned to re-score the branch. In this phase, the final score assigned to the branch is the label that constitutes the majority vote among the three new observers. In the case where there is no majority, the branch is scored as "unknown".

To reduce the number of branches that observers needed to score and thus speed up the scoring process, branches that are very similar to previously scored branches that were labeled as "correct" are accepted automatically. Comparison with previously scored branches is achieved through the use of an intermediate reference, which is the union of all branches up till now that have "correct" as their final label. We use the following two criteria:

1) Centerline overlap: Every point in the centerline is within a 26-neighborhood to a "correct" voxel in the intermediate reference result.
2) Volume overlap: At least 80% of the voxels of the branch are scored as "correct" in the intermediate reference result. Our experiments in a pilot study showed that this threshold of 80% was able to avoid automatically accepting wrong branch segments while not being overly sensitive to small variations.

Branches that fulfill both criteria are automatically scored as "correct" and are exempt from the manual scoring process.

Once all branches from the results of all participating teams are scored, we compute the final reference segmentation for a given image by taking the union of all voxels labeled as "correct" in that image. For the remaining voxels, the voxels that are labeled as "unknown" in the scoring process will be ignored during the evaluation, while the rest are treated as "wrong".

D. Training of human observers

Ten observers took part in the visual scoring. They received a study protocol with scoring instructions, explanations of the software and the different views, and several screenshots of the two views for examples of correct, wrong, and partly wrong segmented branches. Hands-on instruction sessions were set up to further instruct the observers on the evaluation software and scoring procedure. During these sessions, the observers scored at least two complete airway tree segmentations (with the automated branch acceptance option disabled) under the supervision of an experienced observer. The first four observers were trained this way by the first author. The other six observers were trained by one or more of their colleagues, and their agreement with the scores by the experienced observers was computed. When their disagreement with the experienced observers exceeded 10%, an extra session (needed for only 2 out of 10 observers) was held where the errors and correct scores were pointed out.

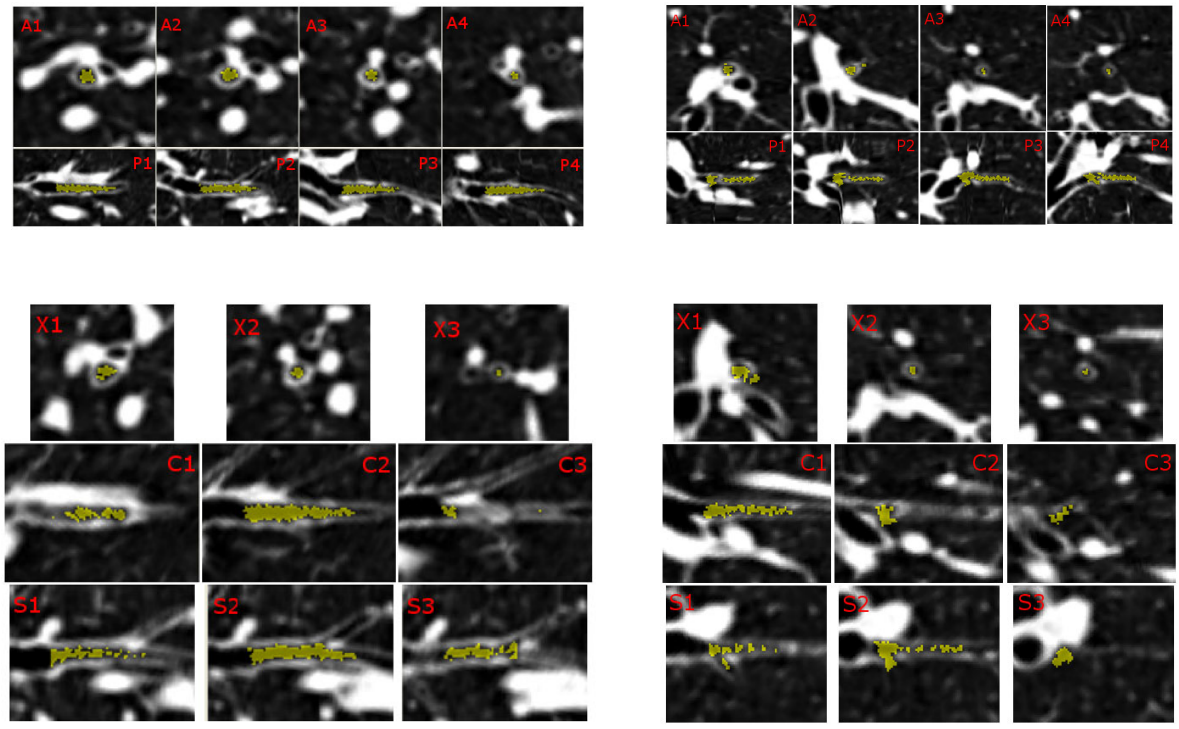


Fig. 3. Example of the reformatted (top panel) and reoriented (bottom panel) views for (a) a correctly extracted branch and (b) a branch with leaks. The alpha numeric characters in the individual images refer to the different cut planes as shown in Figure 2.

IV. Algorithms for Airway Extraction

Ten fully automated algorithms and five semi-automated algorithms (indicated by *) are evaluated in this study. Fully automated algorithms require no manual initialization or interaction and use the same settings for all scans processed. Semi-automated algorithms require user initialization or interaction, which varied from placing a single seed point or selection of certain parameters, to extensive interaction by manually adding or removing complete branches. All evaluated algorithms are briefly described below and an indication of the required processing time per case is provided. All operations are performed in 3D, unless otherwise stated.

1) Morphology based segmentation: Irving et al. [25] use gray scale morphological filtering and reconstruction to detect potential airway regions. The airways are then segmented by a closed space dilation with leakage detection on the marked region. The method takes an average of 71 minutes per image on a 2.83 GHz personal computer (PC).

2) Morphological aggregative: Fetita et al. [26] detect airway candidates using the flood size-drain leveling morphological operator. The airway tree is reconstructed by several propagation schemes applied iteratively to encourage propagation within airways and avoid leakage to the lung parenchyma. Scans are pre-filtered using filter parameters derived from the training set, which are dependent on the scanner model, reconstruction kernel, and dosage. The process takes on average 5 minutes per image.

3) Adaptive cylinder constrained region growing:* Pinho et al. [27] proposed a method to automatically detect the starting point of the trachea and to segment airway branches by applying region growing iteratively within cylindrical volumes of interest. A simplified skeleton constructed based on the starting point and end points of a branch segment is used to estimate the heights, radii and orientations of the cylindrical volumes of interests in the next iteration. A neighbor affinity technique is used to avoid leaks in the region growing algorithm. The method requires specific tuning of the parameter involving the height of the cylindrical volume of interest for certain cases. Segmentation of an image requires less than 8 seconds for most cases on a 2.4 GHz PC.

4) Adaptive region growing and local image enhancement: Feuerstein et al. [28] proposed a tracing scheme that uses cubical volumes constructed based on the orientation and radius of detected branches. The volumes are locally enhanced using a sharpening filter based on a Laplacian of Gaussian kernel. A region growing process is iterated within each of the cubical volumes until a suitable threshold is found, determined by the number of furcating branches. The method takes on average 5 minutes per image on a 2.66 GHz PC.

5) Voxel classification and vessel orientation similarity: Lo et al. [14] perform region growing on the output of a voxel classifier that is trained to differentiate between airway and non-airway voxels. An additional criterion in the region growing allows inclusion of lower probability airway candidates if their orientation is sufficiently similar to that of a nearby blood vessel, exploiting the fact that airways and arteries run parallel to each other. The framework takes approximately 90 minutes per image on a 2.66 GHz PC.

6) Two-pass region growing and morphological gradient:* Fabijańska [29] proposed a two step segmentation approach.

The first step consists of obtaining an initial segmentation by performing region growing on an image where the intensities are normalized. The initial segmentation is then used as seeds for a second region growing process that is performed on the morphological gradient of the original image. The method requires manual selection of a threshold related to the second region growing in some cases. Computation time is less than 10 minutes for a typical chest CT on a 1.66 GHz PC.

7) Tube detection and linking: Bauer et al. [30] proposed a method to reconstruct the airway tree from detected airway branches. Therefore, they utilize a tube detection filter with a ridge traversal procedure to extract centerlines of dark tubular structures in the CT image. The airway tree is reconstructed starting from the trachea by iteratively connecting these tubular structures. During this process, prior knowledge about the structure of the airway tree, such a branching angle and radius, is used. Segmentation of a single dataset takes on average 3 minutes using a graphics processing unit (GPU) based implementation of the tube detection filter.

8) Maximal contrast adaptive region growing:* Mendoza et al. [31] use region growing with maximal-contrast stopping criteria. Local non-linear normalization using a sigmoidal transfer function and denoising via an in-slice bidimensional median filter are introduced to improve robustness. The method requires the user to manually initialize several seeds in the trachea region so that the statistical nature of air density values can be characterized for each case. Segmentation of a single case requires on average 2 minutes on a 2 GHz PC.

9) Centricity-based region growing: Wiemker et al. [32] proposed a voxel-wise centricity measure in combination with prioritized region growing. The centricity measure quantifies how central a given voxel is to the surrounding airway walls by measuring the lengths of rays cast isotropically in 3 dimensions. A ray terminates if the intensity difference, with respect to the starting point, of a point along a ray is higher than a certain threshold. A region growing process is used to obtain the actual segmentation, where it proceeds until all connected voxels below a certain intensity threshold and above a certain minimum centricity value are extracted. The runtime of the method for an image is 19 seconds on average on a 3 GHz PC.

10) Adaptive region growing within local cylindrical volumes of interest: Lee et al. [33] proposed another local adaptive region growing method. To avoid leaks, the region growing is performed within local volumes of interest and requires that at least half of the neighbors of a candidate voxel are below a certain threshold. The threshold is incremented until leaks are detected. Segmentation of an image takes less than 30 seconds on a 3 GHz PC.

11) Template matching:* Born et al. [34] proposed 2D template matching technique and a set of fuzzy rules to detect and prevent leakage. Airway tree segmentation is obtained through an iterative procedure that iterates between 3D region growing, 2D wave propagation and 2D template matching. Their method requires the user to set a seed point in the trachea manually. The method takes around 25 seconds per image on a 2.4 GHz PC.

12) Adaptive region growing with histogram correction: Weinheimer et al. [35] proposed an adaptive region growing approach that monitors the volume of the segmented region, and increases the threshold if no leakage is detected. The acceptance criteria in the region growing process are based on fuzzy logic rules and on rays cast from the voxel in the axial, coronal, and sagittal plane. Histogram analysis is used to preprocess the CT scan and to dynamically adapt the fuzzy logic rules based criteria to different images. An average of 3 minutes is required to segment a case on a 2.83 GHz PC.

13) Gradient vector flow: Bauer(a) et al. [36] proposed a method utilizing properties of the Gradient Vector Flow (GVF) [43] vector field. A measure of tube-likeness is computed for every voxel based on the vector field obtained from the GVF. Subsequently, the airway tree centerlines are extracted by applying hysteresis thresholding on the tube-likeness map. The final segmentation is obtained by following the gradient flow path in the inverse direction and adding the voxels along the path until maximum gradient magnitude is reached. Using a GPU based implementation of the GVF, the method requires 6 minutes to process a dataset.

14) Multi-threshold region growing: Van Rikxoort et al. [37] proposed a wavefront propagation approached that is based on sphere constricted region growing, where geometric characteristics of a branch such as furcation and radius are obtained from the propagating front. A series of rules, such as radius growth, furcating angles, etc., are used to detect and prevent leaks. The method also features a multi-threshold approach, where the threshold used is increased as long as no leaks are detected. Segmentation of an image takes around 10 seconds on a single-core PC.

15) Automated region growing with manual branch adding and leak trimming:* Tschirren et al. [38] proposed an interactive segmentation tool. An initial airway tree segmentation obtained with a region growing method that uses an optimal threshold selected based on the volume of the extracted region. The tree is subdivided in branches by skeletonization. The user can manually select leaks to remove and add new branches by placing seed points. The new branches are formed using region growing and connected to the initial segmentation using the Dijkstra algorithm. An average of 59 minutes of human interaction time is required to segment a single image.

To assess whether the different algorithms provide complementary information and whether results can be improved by combining algorithms, we evaluate additional segmentations that combine segmentation results from several algorithms. A voxel based fusion scheme is used for this purpose, in which a voxel is labeled as part of the airway tree if it is marked as airways by at least T_f algorithms.

We use sequential forward selection (SFS) to select which algorithms to include in the fusion schemes. The SFS procedure starts with the algorithm that produced the maximum total tree length, and at each subsequent iteration adds the algorithm that gives the largest increase in the total tree length obtained by the combined segmentations. In addition, we investigated fusion schemes including only the fully automatic algorithms, as well as algorithm selection based on computation time.

To appear in IEEE Transactions on Medical Imaging; http://dx.doi.org/10.1109/TMI.2012.2209674

V. Evaluation metrics

In order to compare the results of the different algorithms in a standardized manner, centerlines are first computed for all segmentation results and for the reference, using the algorithms described in Section IV. To determine the length of a branch in a given segmentation, we compute the length of the centerline of that branch after projection to the reference segmentation centerline. In this way, a bias due to, for example, high tortuosity in the supplied centerline, is avoided. Branches are counted as "detected" by the segmentation results of an algorithm if they are at least $\Delta l = 1$ mm long.

Three performance measures are computed for each segmented airway tree:

1) Branches detected: The percentage of branches that are detected correctly with respect to the total number of branches present in the reference, N_{ref}, defined as
$$\frac{N_{seg}}{N_{ref}} \times 100\%$$
where N_{seg} is the number of branches detected correctly by the segmentation.
2) Tree length detected: The fraction of tree length that is detected correctly relative to the total tree length in the reference, L_{ref}, defined as
$$\frac{L_{seg}}{L_{ref}} \times 100\%$$
where L_{seg} is the total length of all branches detected by the segmentation.
3) False positive rate: The fraction of the segmented voxels that is not marked as "correct" in the reference, defined as
$$\frac{N_w}{N_c + N_w} \times 100\%$$
where N_c and N_w are the number of voxels in the segmented airway that overlap with the "correct" and "wrong" regions in the reference respectively. Note that "unknown" regions in the reference are not included in the calculation of the false positive rate.

The trachea is excluded from all measures. Further, for measure 3, the left and right main bronchi are excluded as well.

VI. Results

A. Observer agreement

A total of 40,772 branches were evaluated. Among these, 52.16% were accepted automatically, 33.16% were assigned a final score at phase 1, and 14.67% were assigned a final score at phase 2. Of the branches, 82.59% were scored as "correct", 10.77% were scored as "wrong", 5.51% were scored as "partly wrong" and 1.13% were scored as "unknown".

The final reference segmentation contained 81.02% voxels labeled as correct, 11.16% as "partly wrong", 7.12% as "wrong", and 0.70% as "unknown", where the trachea and the left and right main bronchi were excluded when computing the percentages. We found that most voxels inside the airway lumen that were originally part of a "partly wrong" branch were detected correctly by one of the other algorithms and were relabeled as "correct". We therefore counted the remaining "partly wrong" voxels as "wrong", while all "unknown" voxels were ignored in the evaluation.

TABLE II
Confusion matrix of observer scores, where the columns indicate the scores assigned by the observers and the rows indicate the final scores used to construct the reference.

		Observer			
		Correct	Partly wrong	Wrong	Unknown
Final	Correct	23,666	1,626	364	47
	Partly wrong	1,885	6,319	685	30
	Wrong	2,723	1,843	13,135	476
	Unknown	768	657	764	87

TABLE III
The number of scores # from all observers and average agreement across observers for branches of different sizes, measured in number of voxels.

Size	#	Mean agreement(%)
≤ 200	36,385	77.88
>200 & ≤ 400	8,421	82.90
>400 & ≤ 600	4,264	86.78
>600 & ≤ 800	2,249	85.83
>800 & ≤ 1000	1,269	87.46
>1000 & ≤ 1200	783	87.71
>1200 & ≤ 1400	486	87.13
>1400	1,218	92.02

Table II presents the confusion matrix of the 55,075 individual scores given by the observers (in both phases of the scoring process) in comparison to the final scores for each branch. The average percentage of assigned scores that were in agreement with the final scores was 80.31%, with a standard deviation of 10.68%. In this computation, observers are counted as in agreement irrespective of their original score if the final score is "unknown". The majority of disagreement is between the labels "partly wrong" and "correct" or "wrong"; in 5.6% of cases, there is disagreement whether a branch is "correct" or "wrong". Table III presents the average agreement between the scores from the observers and the final scores for branches of different sizes, where the size is given in terms of number of voxels.

B. Completeness of the established reference

The reference standard in this work is based on visual assessment of the correctness of airway branches produced by any of the participating algorithms. Therefore it does not include airways that were missed by all algorithms, and thus the two sensitivity measures reported in this study, branches detected and tree length detected, have been overestimated. In order to provide a rough estimate of the number of missing branches in the reference standard, an additional observer study was conducted. A trained human observer inspected 200 random axial slices from the 20 test scans. For these slices, the lung masks and the overlay with the reference airway segmentation could be toggled on and off, and inspection in 3D with coronal and sagittal views was available. The observer clicked every point that he deemed could represent a missed airway branch, inspected the three orthogonal views and scrolled through the axial slices and decided if this was indeed a missed branch. If the branch bifurcated and child

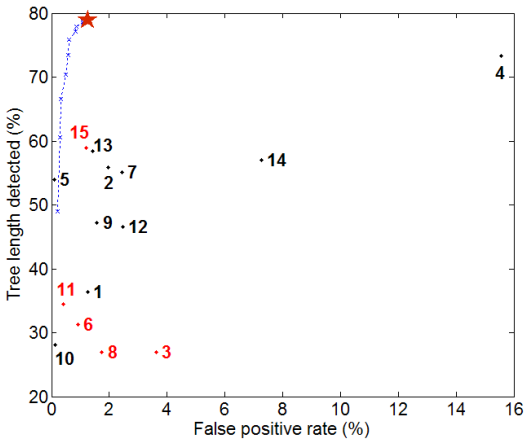


Fig. 4. Average tree length versus average false positive rate of all algorithms, with the algorithms in the semi-automated category in red. The fusion scheme ($T_f = 2$) combining all 15 algorithms is indicated with $\star$. The blue line indicates the fusion results of including different number of algorithms, starting from 2 to all 15 algorithms.

branches were visible in the same slice, these child branches were indicated as well.

The reference standard contained on average 247.9 branches per scan. The observer added on average 0.56 airways per slice. From the reference standard, we computed that each of the terminal branches is visible in 9.6 slices on average. The test scans contained on average 431 slices. From these numbers we can compute that on average 25.1 branches were missed per scan ($0.56 \times 431 / 9.6$), and this is around 10% of all branches in the reference.

C. Comparison of algorithms

Table IV(A) presents the three evaluation measures for the 15 algorithms. The evaluation measures for the fusion scheme with SFS procedure are given in Table IV(B). Figure 4 gives an overview of the average performance of the different algorithms using a scatter plot of tree length detected versus false positive rate. Figure 5 shows box plots of tree length and false positive rate for the different algorithms. Box plots in Figure 6 give the number of correctly detected branches and the volume of the "wrong" voxels, or leakage volume, per case.

In the box plots, the red line indicates the median, and the lower and upper edge of the box indicate the 25th and 75th percentile respectively. The lines below and above the box, or "whiskers", represent the largest and smallest values that are within 1.5 times the interquartile range, while the red open circles show all outliers outside this range.

Surface renderings of two cases are given in Figure 7 and Figure 8, with correct and wrong regions indicated in green and red respectively.

D. Local sensitivity analysis

We evaluated the detection rate of different anatomical branches for all algorithms. Anatomical branch labels were assigned manually in the reference airway trees down to the segmental bronchi using the Pulmonary Workstation software

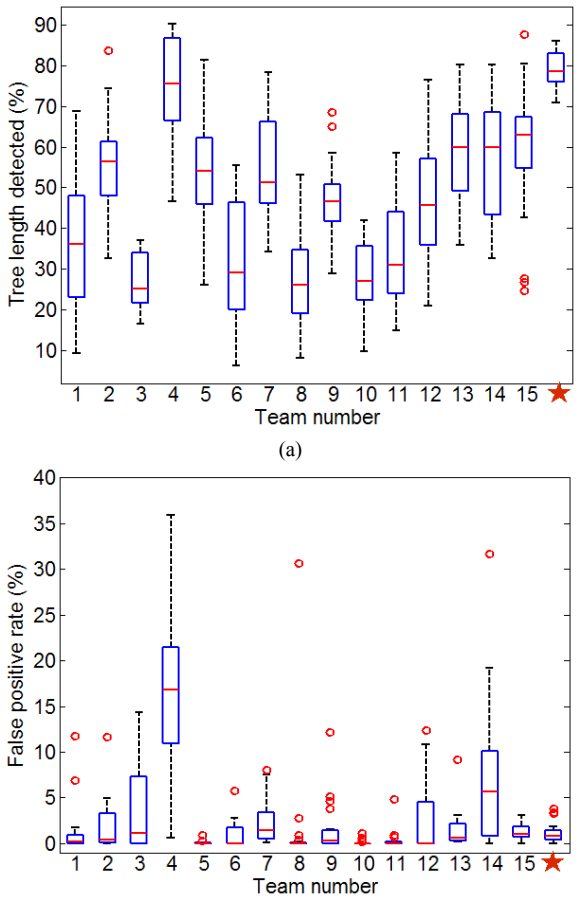


Fig. 5. Box plots of (a) tree length and (b) false positive rate of the algorithms. The fusion scheme combining all 15 algorithms is indicated with $\star$.

package (VIDA Diagnostics, Coralville, Iowa, USA). Figure 9a shows a surface rendering of the manually labeled reference airway tree, with the different anatomical labels shown using different colors. For each labeled branch in the reference, we determine whether an algorithm detects the branch by comparing branch centerlines. Figure 9b presents a diagram showing the sensitivity of the algorithms to the different anatomical labeled branches, which is def ned as the number of algorithms that detected (part of) a branch by the total number of algorithms. A scatter plot of the average sensitivity for the lobar and segmental branches for the individual algorithms is given in Figure 10.

E. Combination of algorithms

Figure 13 shows a bar plot of the percentage of branches detected versus the number of algorithms that detected them, averaged across all test cases. Figure 12 shows the surface renderings of the reference segmentations of the test set, with the branches color coded according to the number of algorithms that detected them. More than 30% of the branches were on average extracted by three algorithms or less. A fusion scheme that combines results from all participating algorithms, as proposed in Section VI-E, was able to extract more complete airway trees than any of the individual algorithms, as can be seen in Table IV and Figure 4. The results from the fusion

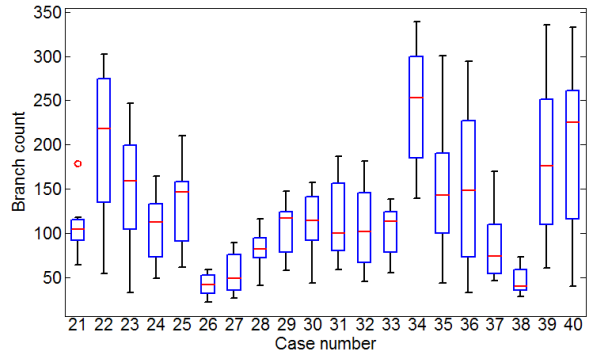
TABLE IV

(A) Evaluation measures averaged across the 20 test cases of each algorithm. * indicates teams in the semi-automated category.

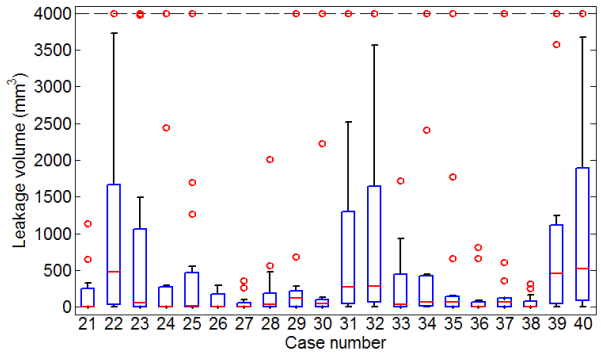
	Branches detected (%)	Tree length detected (%)	False positive rate (%)
(1) Irving *et al.*	43.5	36.4	1.27
(2) Fetita *et al.*	62.8	55.9	1.96
(3*) Pinho *et al.*	32.1	26.9	3.63
(4) Feuerstein *et al.*	76.5	73.3	15.56
(5) Lo *et al.*	59.8	54.0	0.11
(6*) Fabijańska	36.7	31.3	0.92
(7) Bauer *et al.*	57.9	55.2	2.44
(8*) Mendoza *et al.*	30.9	26.9	1.75
(9) Wiemker *et al.*	56.0	47.1	1.58
(10) Lee *et al.*	32.4	28.1	0.11
(11*) Born *et al.*	41.7	34.5	0.41
(12) Weinheimer *et al.*	53.8	46.6	2.47
(13) Bauer(a) *et al.*	63.0	58.4	1.44
(14) van Rikxoort *et al.*	67.2	57.0	7.27
(15*) Tschirren *et al.*	63.1	58.9	1.19
Fusion of 15 algorithms ($T_f = 2$)	84.3	78.8	1.22

(B) Evaluation measures, averaged across the test cases, of the fusion scheme using different number of algorithms with $T_f = 2$. The actual algorithms, selected by SFS, are indicated in the brackets beside the number of algorithms used.

number of algorithms	Branch detected (%)	Tree length detected (%)	False positive rate (%)
2 (4 & 2)	56.2	49.0	0.22
3 (+15)	67.1	60.6	0.29
4 (+13)	72.9	66.6	0.33
5 (+14)	77.3	70.4	0.49
6 (+7)	79.0	73.4	0.58
7 (+5)	80.9	75.9	0.60
8 (+12)	82.2	77.1	0.84
9 (+1)	83.1	77.8	0.86
10 (+9)	83.9	78.4	1.08
11 (+11)	84.3	78.6	1.08
12 (+6)	84.1	78.8	1.14
13 (+8)	84.2	78.8	1.15
14 (+10)	84.2	78.8	1.16
15 (+3)	84.3	78.8	1.22



(a)



(b)

Fig. 6. Box plots of (a) branch count and (b) leakage volume, with the maximum leakage volume clipped at 4000 mm^3, of the 20 test cases computed across the 15 participating algorithms.

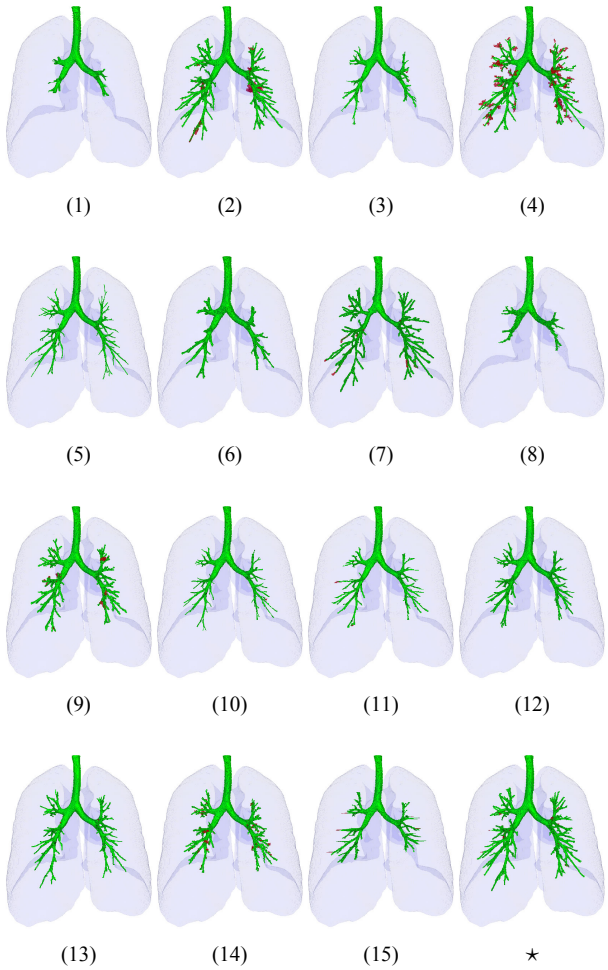


Fig. 7. Surface renderings of results for case 23, with correct and wrong regions shown in green and red respectively.

scheme have the highest average tree length with a relatively low average false positive rate. The blue line in Figure 4 shows the improvement in performance of the fusion scheme with an increasing number of algorithms included in the selection.

A T_f of two was used for the final fusion scheme shown in Table IV and Figure 4. It was observed that although further increasing T_f reduces the false positive rate, the tree length

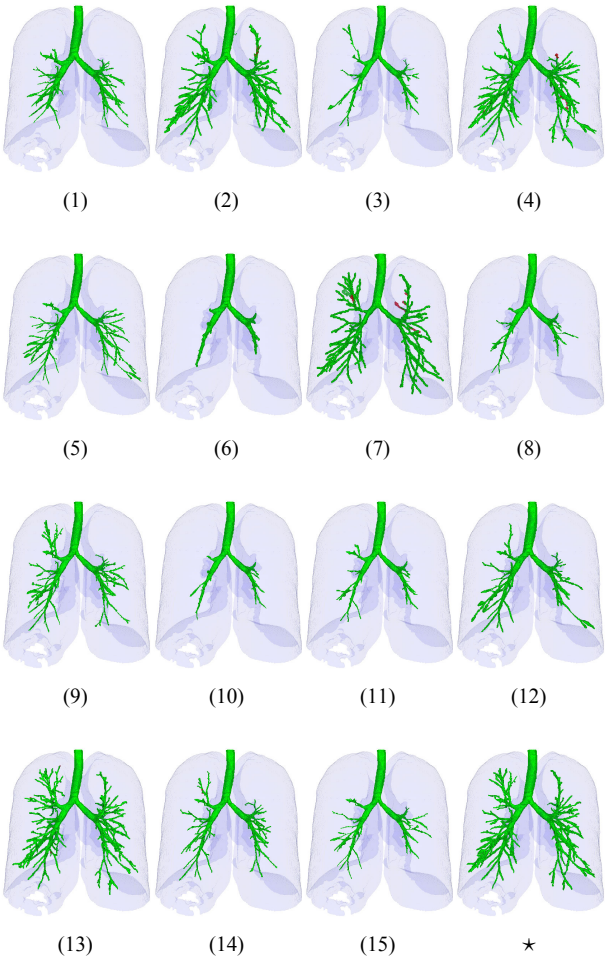


Fig. 8. Surface renderings of results for case 36, with correct and wrong regions shown in green and red respectively.

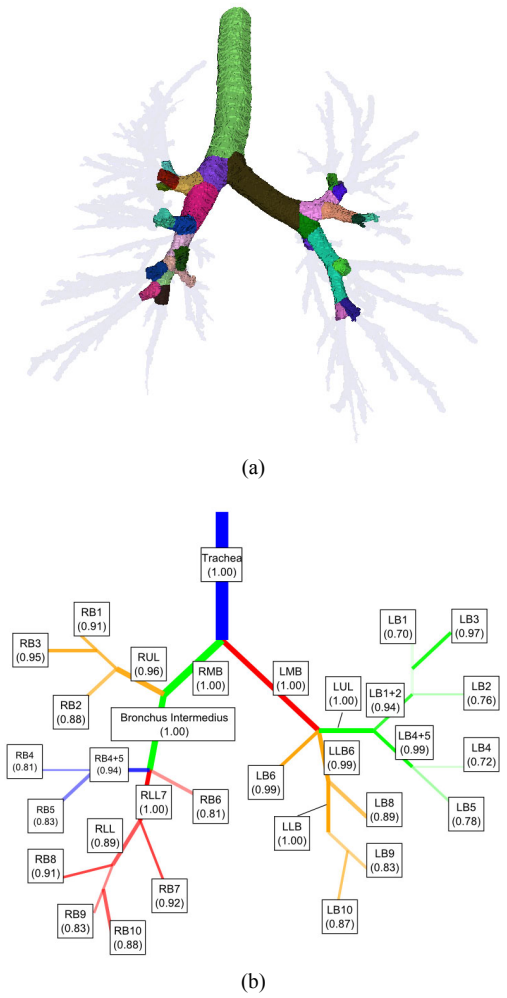


Fig. 9. (a) An example of a reference airway segmentation with the manually assigned anatomical labels, where the different colors indicate different anatomical labels. (b) Branch detection sensitivity for different labeled branches averaged over all cases.

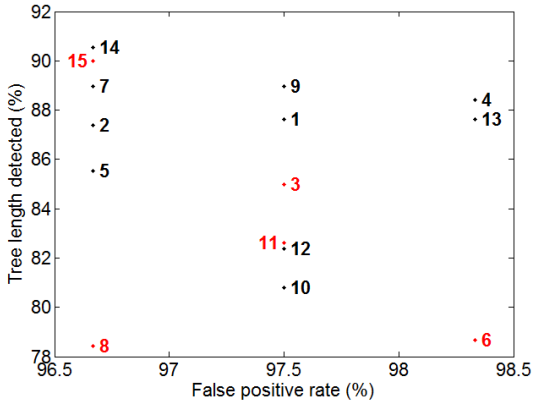


Fig. 10. Scatter plot of the sensitivity, averaged over all cases, of the lobar and segmental branches for the 15 algorithms.

detected and branches detected were greatly reduced as well. For example, with $T_f = 3$, the resulting false positive rate, tree length detected and branches detected were 0.14%, 66.4% and 74% respectively. False positive rate was observed to drop to 0% at $T_f = 7$, with 44.3% of tree length detected and 53.1% of branches detected.

The results of the fusion scheme in Table IV(B) consist of algorithms from both the automated and the semi-automated category. To investigate the usage of the fusion scheme in a more practical setting, we performed additional experiments using the sequence from SFS from Table IV(B) with semi-automated algorithms excluded. We also investigate the effects of incrementally fusing from the least to the most computationally intensive algorithms, based on the reported average execution time required per case, in the following sequence: algorithm 14, 9, 10, 7, 12, 4, 2, 13, 1 and 5. Figure 11 shows a graph of tree length detected against false positive rate of the fusion scheme with increasing number of algorithms included, for the sequence obtained from SFS and based on the reported execution time.

VII. Discussions

A. Performance of different algorithms

Fifteen algorithms for airway extraction have been compared in this study. Performance varies widely, as is most

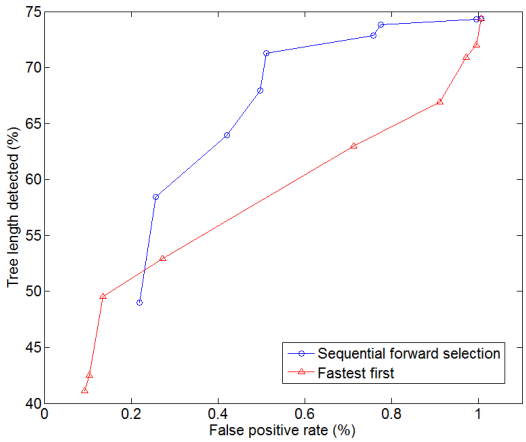


Fig. 11. Tree length detected and false positive rate for an increasing number of fully automated algorithms included in the fusion scheme, with sequence based on SFS and on execution time from fastest to slowest.

obvious from the renderings in Figures 7 and 8. There is a clear tradeoff between sensitivity and specif city in the airway tree extracted by the different algorithms. This is shown in Figure 4 and Figure 5, where it is observed that more complete trees are often accompanied by more false positives. The most conservative algorithm, algorithm 10, obtains the smallest average false positive rate (0.1%) and is also among the algorithms with the lowest average tree length (32.4%). On the other hand, algorithm 4 is the most explorative algorithm, yielding the highest average tree length (76.5%), but at the expense of the highest average false positive rate (15.6%).

In general, semi-automatic algorithms perform no better than fully automatic algorithms. This is probably due to the fact that manual interactions for semi-automatic algorithms are limited to selecting initial seed points for the trachea (algorithm 8 and 11) or tuning parameters manually (algorithm 3 and 6) for a few test cases. The only algorithm with extensive interaction is algorithm 15, where branches could be added or removed by users until they were satisf ed with the f nal segmentation result. Despite the interaction time of on average one hour per case, the overall results for the performance metrics used in this study are close to those of algorithm 13, which is fully automatic.

Because of the use of different types of CPUs and, in some cases, GPUs during execution, it is not possible to directly compare the execution time of the different algorithms. However, we do observe a wide range of execution time, from less than thirty seconds to more than one hour. Most execution times are between two to f ve minutes per case.

Interestingly, no algorithm comes close to detecting the entire reference airway tree, as observed from Figure 4. The highest branches detected and tree length detected for each case ranges from 64.6% to 94.3% and 62.6% to 90.4%, respectively, with an average branches detected and tree length detected of less than 77% and 74%, respectively. Fusing results from the participating algorithms improves the overall result substantially, reaching an average number of branches detected of 84.3% and an average tree length detected of 78.8%, with an average false positive rate of only 1.22%, when all f fteen

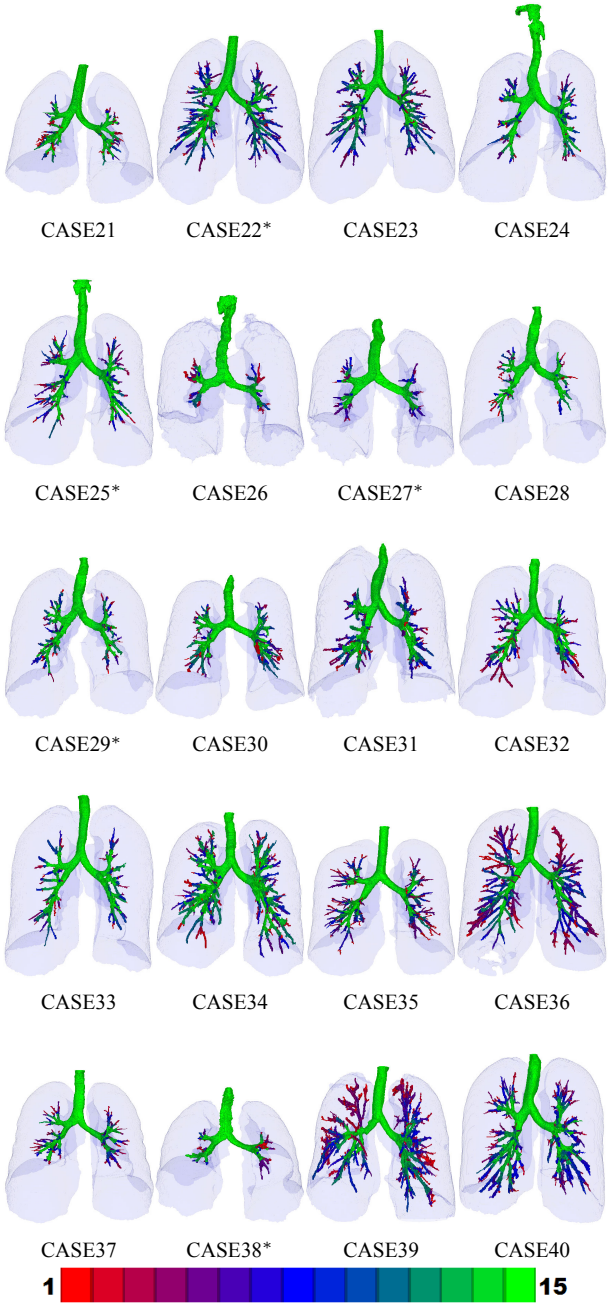


Fig. 12. Surface renderings of the reference. * indicates that the case is from the same subject as the preceding case. The branches are color coded from red (detected by a single algorithm) to green (detected by all 15 algorithms).

algorithms were used.

Experiments on the inclusion of the results from different algorithms using the SFS procedure show that the tree length of the fused results converges quite rapidly, as displayed in Figure 4 and Table IV(B). This indicates that reasonably good results can be obtained by fusing only a subset of the algorithms. In fact, Table IV(B) shows that with a smaller number of algorithms (e.g. using up to 9 algorithms) in the fusion procedure, one can obtain a lower false positive rate at almost the same sensitivity.

The performance of the fusion scheme degraded slightly, to a tree length detected of 74.3% and a false positive

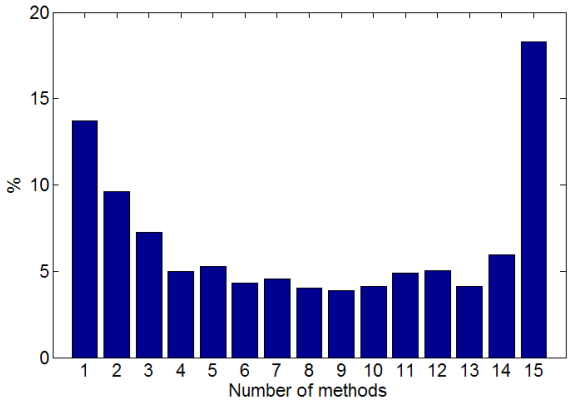


Fig. 13. Bar plot shows the percentage of branches detected vs. the number of algorithms that detected them, averaged across all test cases. Figure shows 18.3% of the branches were detected by all 15 methods, while 13.7% of the branches were only detected by one algorithm.

rate of 1.01%, when only fully automated algorithms were included, with an approximate cumulative execution time of 184 minutes. Despite the drop in performance, the tree length detected is still higher than that of any of the individual algorithms. As expected, performance of the fusion scheme using the sequence from SFS converges more rapidly than simply ordering the algorithms based on their execution time. Using the the sequence from SFS, the fusion scheme reaches a tree length detected of more than 70%, or 71.3% to be exact, with only six algorithms. Although the sequence ordered according to execution time requires eight algorithms in the fusion scheme to reach a tree length detected of 70.9%, it does have lower cumulative computation time (approximately 23 minutes per case) as compared to the sequence from SFS (approximately 109 minutes per case).

As the results from the fusion scheme are derived from the same segmentations that were used to construct the reference standard, performance of the combined algorithm may be slightly lower on unseen data. However, the fact that the results from the fusion scheme are better than those of individual algorithms indicates that the different extraction algorithms are complementary to each other and their combination can be expected to improve results. Such a property is not unique and has been noted in other comparative studies [44], [45].

Figure 9b shows that the algorithms have fairly high sensitivity in detecting the segmental bronchi, ranging from 0.70 (LB1) to 0.99 (LB6), indicating that each of the segmental bronchi is at least detected by 10 different algorithms on average.

B. Reference standard

This work has presented a novel way to construct a reference standard for a structure that is hard to segment manually, in this case the airway tree, from multiple machine made segmentations. The key concept is to break the machine made segmentation into parts, which is a natural operation for airway trees as they consist of branches, and have human experts accept or reject the parts. Overall, this procedure, though time-consuming, worked well and has resulted in a unique resource, a reference standard that is available to the research community for algorithm evaluation.

A limitation of our reference standard is that it, by the nature of the way in which it was constructed, does not contain all visible airway branches in the data set. Even though one of the algorithms (Algorithm 15) employed extensive user interaction of up to three hours per scan, there are visible airways that have not been indicated by any of the 15 algorithms. We therefore conducted an additional study, described in Section VI-B, from which we concluded that about 10% more visible branches are presented in the data. Although it has to be realized that this is an estimate only, based on the opinion of a single human observer who has to make subjective judgements about the visibility of very small airways, we can conclude that the reported sensitivities from the algorithms in this study have a positive bias. If in the future new results were submitted and processed in a similar manner, by having human observers assess the correctness of new branches, it is possible that this percentage of missed airways would decrease somewhat.

A point of concern on the credibility of the reference standard would be the relatively low overall agreement between the scores from the observers and the final scores, which averaged to 80.31% across observers. This low overall agreement is mainly caused by the small branch segments, for which it is often difficult to discern whether they are true airway branches or not. The observers in our study have especially low agreement (of less than 85%) with the final scores for branch segments smaller than 400 voxels, as shown in Table III. Although the scores of these small branch segments of less than 400 voxels constitute 81.35% of the overall scores assigned by the observers, they only consist of 20.75% in terms of volume.

Another limitation of our approach is that we take the union of "correct" voxels as the reference airway tree and as a result the correct part of voxels in branches labeled as "partly wrong" will be treated as wrong and be penalized during the evaluation. However, as segmented airway trees from different algorithms of the same scan are used, most of the voxels that are previously marked as "partly wrong" will eventually be assigned different labels, as they overlap with either "correct" or "wrong" regions of branches from other algorithms. Although there are still correct voxels within "partly wrong" regions being discarded, the impact on the evaluation results is minimal as it concerns only a small fraction of the original 5.51% of "partly wrong" voxels.

C. Case analysis

The dataset used in this study is designed to evaluate performance of airway extraction algorithms over a wide range of different variations and anomalies. It is not a suitable dataset for the study of effects of specific factors, such as dose, inspiration level, pathology etc., have on the performance of airway extraction algorithms, due to the small number of cases used and the fact that each case had multiple confounding factors that may be influencing the results. Nevertheless, we investigated the effects of the different factors based on the

small amount of paired scans and groups of scans with similar characteristics in our dataset.

To study the effect of the different doses, we separated the scans into three groups based on their tube voltage and tube current: a low dose group (cases 24, 26 and 38, with a mean branches detected of 51.9% and mean false positive rate of 1.30%), an intermediate dose group (cases 21-23, 25, 27, 30-33, 37, 39 and 40, with a mean branches detected of 51.5% and mean false positive rate of 3.54%) and a diagnostic dose group (cases 28, 29, and 34-36, with a mean branches detected of 52.6% and mean false positive rate of 1.96%). Using unbalanced one-way Analysis of Variance (ANOVA), no significant difference ($p = 0.92$) in branches detected were found between the three groups. However, we did find significant difference in false positive rate ($p = 0.02$) between the groups, where significant difference were detected between the intermediate and low dose group ($p = 0.02$), and between intermediate and diagnostic dose group ($p = 0.03$), but not between the low and diagnostic dose group ($p = 0.37$). For the two pairs of low dose and intermediate dose scans (cases 24 and 25, and 26 and 27), branch counts were significantly lower ($p < 0.01$ from paired Student's t-tests) for the low dose scans (mean branch count of 73.4 and mean leakage volume of 322.3 mm^3) than the intermediate dose scans (mean branch count of 92.9 and mean leakage volume of 376.8 mm^3), while there was no significant difference in leakage volume ($p = 0.54$).

From the available paired inspiration and expiration scans (cases 21 and 22, and 37 and 38), not only did the segmentations of the inspiration scans have more correct branches, they also had more leakage than their expiration counterparts. Inspiration scans exhibited an average branch count of 145 branches and leakage volume of 942 mm^3 compared to 76 branches and 115 mm^3 for expiration scans. A paired Student's t-tests showed that these difference were significant ($p < 0.01$ for branch count and $p = 0.02$ for leakage volume). It should be noted however, that scan 38 was acquired with a lower dose and a different reconstruction kernel, which could have affected the results as well.

The image pair case 28 and case 29 consists of scans from the same subject reconstructed with a soft and a hard kernel, respectively. Significantly more branches ($p < 0.01$) were extracted from the scan constructed using the hard kernel, with an average of 106 branches compared to 80 branches from the soft kernel reconstructed scan. The average leakage volume for the hard kernel scan was higher, 418 mm^3 compared to 236 mm^3, but the difference was not significant ($p = 0.30$).

The different noise levels from Table I, low (mean branches detected of 51.2% and mean false positive rate of 2.65%), middle (mean branches detected of 52.3% and mean false positive rate of 2.32%) and high (mean branches detected of 52.1% and mean false positive rate of 3.56%), did not seem to have much effect on either the branches detected or false positive rate, with a p-value 0.90 and 0.30 respectively via unbalanced one-way ANOVA. For the scans that were classified visually as middle (mean branches detected of 53.0% and mean false positive rate of 2.43%) and hard (mean branches detected of 51.5% and mean false positive rate of 3.80%) reconstruction (the soft reconstruction group was left out as it only had a single case), although no difference was found on the branches detected ($p = 0.52$), we did find a slight difference in the false positive rates ($p = 0.0549$).

Additionally, we also performed unpaired Student's t-tests on group of scans without obvious abnormalities (cases 21, 22, 23, 28, 29, 35 and 37) and a group of scans showing bronchiectasis (cases 33, 34, 36 and 39). Mean branches detected and mean false positive rate were 53.8% and 2.75% for the healthy group, and 48.1% and 2.63% for the bronchiectasis group. We did not find a significant difference in branches detected ($p = 0.08$) and false positive rate ($p = 0.89$) between both groups.

D. Future of EXACT

All training and test data are publicly available at the EXACT'09 website[3]. This website also provides detailed descriptions for each algorithms, the performance metrics for each scan and each algorithm, and surface renderings for the results from each algorithm for all test cases. We also provide the opportunity to have new results evaluated against the current reference standard. The downside of this is that some correctly segmented branches from newly submitted algorithms may be classified as incorrect if they are missing from the current reference standard. To solve this, we hope to organize a future round of human observer evaluation where the reference tree will be updated with additional branches found by the newly submitted results and previously submitted results will be re-evaluated.

VIII. Conclusion

A framework has been presented to establish a reference airway tree segmentation. This was used to evaluate airway extraction algorithms in a standardized manner. This is the first study that performed quantitative evaluation of a large number of different airway tree extraction algorithms (a total of fifteen algorithms), which were applied to a single dataset (twenty chest CT scans from various institutes) and evaluated in a common, fair, and meaningful way. Three performance measures were used to evaluate the sensitivity and specificity of the different algorithms. Results showed that no algorithm was capable of extracting more than an average of 74% (range 62.6% to 90.4%) of the total length of all branches in the reference, with an average false positives of 2.81% (range 0.11% to 15.56%). It was shown that better results can be obtained by a simple fusion scheme that retains regions that are marked by two or more algorithms, resulting in extracting on average 78.84% of the total length of all branches in the reference, with an average false positive rate of only 1.22%.

Acknowledgment

This work was funded in part by the Danish Council for Strategic Research (NABIIT), the Netherlands Organization for Scientific Research (NWO), and by grants HL080285 and HL079406 from the U.S. National Institutes of Health.

[3]See http://image.diku.dk/exact/

The authors would like to thank the following people for their help in providing the scans for the study:

- Haseem Ashraf and Asger Dirksen (Gentofte University Hospital, Denmark)
- Patrik Rogalla (Charité, Humboldt University Berlin, Germany, now at University of Toronto, Canada)
- Jan-Martin Kuhnigk (Fraunhofer MEVIS, Germany)
- Berthold Wein (University Hospital of Aachen, Germany)
- Atilla Kiraly and Carol Novak (Siemens Corporate Research, USA)

They would like to also thank the following for participating in the study:

- Françoise Prêteux (Institut TELECOM / Telecom Sud-Paris, France)
- Pierre-Yves Brillet (Avicenne Hospital, France)
- Philippe Grenier (Pitié-Salpêtrière Hospital, France)
- Paul Taylor and Andrew Todd-Pokropek (University College London, UK)
- Sten Luyckx (University of Antwerp, Belgium)
- Takayuki Kitasaka (Nagoya University, Japan)
- Jon Sporring (University of Copenhagen, Denmark)
- Thomas Pock and Horst Bischof (Graz University of Technology, Austria)
- Begoña Acha and Carmen Serrano (Universidad de Sevilla, Spain)
- Thomas Bülow and Cristian Lorenz (Philips Research Lab Hamburg, Germany)
- Dirk Iwamaru (CADMEI GmbH, Ingelheim, Germany)
- Matthias Pfeife (University Hospital Tübingen, Germany)
- Dirk Bartz (Universität Leipzig, Germany)
- Tobias Achenbach and Christoph Düber (University of Mainz, Germany)
- Wouter Baggerman (University Medical Center Utrecht, The Netherlands)

REFERENCES

[1] Y. Nakano, S. Muro, H. Sakai, T. Hirai, K. Chin, M. Tsukino, K. Nishimura, H. Itoh, P. D. Paré, J. C. Hogg, and M. Mishima, "Computed tomographic measurements of airway dimensions and emphysema in smokers. Correlation with lung function," *American Journal of Respiratory and Critical Care Medicine*, vol. 162, no. 3 Pt 1, pp. 1102–1108, Sep 2000.

[2] P. Berger, V. Perot, P. Desbarats, J. M. T. de Lara, R. Marthan, and F. Laurent, "Airway wall thickness in cigarette smokers: quantitative thin-section CT assessment," *Radiology*, vol. 235, no. 3, pp. 1055–1064, Jun 2005.

[3] S. Ukil, M. Sonka, and J. M. Reinhardt, "Automatic segmentation of pulmonary fssures in X-ray CT images using anatomic guidance," in *Medical Imaging 2006: Image Processing*, J. M. Reinhardt and J. P. W. Pluim, Eds., vol. 6144, no. 1. SPIE, 2006, p. 61440N.

[4] X. Zhou, T. Hayashi, T. Hara, H. Fujita, R. Yokoyama, T. Kiryu, and H. Hoshi, "Automatic segmentation and recognition of anatomical lung structures from high-resolution chest CT images," *Computerized Medical Imaging and Graphics*, vol. 30, no. 5, pp. 299–313, July 2006.

[5] K. Mori, Y. Nakada, T. Kitasaka, Y. Suenaga, H. Takabatake, M. Mori, and H. Natori, "Lung lobe and segmental lobe extraction from 3D chest CT datasets based on fgure decomposition and Voronoi division," in *Medical Imaging 2008: Image Processing*, J. M. Reinhardt and J. P. W. Pluim, Eds., vol. 6914, no. 1. SPIE, 2008, p. 69144K.

[6] E. M. van Rikxoort, M. Prokop, B. de Hoop, M. A. Viergever, J. Pluim, and B. van Ginneken, "Automatic segmentation of pulmonary lobes robust against incomplete fssures," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1286–1296, 2010.

[7] K. S. Lee and P. M. Boiselle, "Update on multidetector computed tomography imaging of the airways." *Journal of Thoracic Imaging*, vol. 25, no. 2, pp. 112–124, May 2010.

[8] M. W. Graham, J. D. Gibbs, D. C. Cornish, and W. E. Higgins, "Robust 3-D airway tree segmentation for image-guided peripheral bronchoscopy." *IEEE Transactions on Medical Imaging*, vol. 29, no. 4, pp. 982–997, Apr 2010.

[9] A. P. Kiraly, W. E. Higgins, G. McLennan, E. A. Hoffman, and J. M. Reinhardt, "Three-dimensional human airway segmentation methods for clinical virtual bronchoscopy," *Academic Radiology*, vol. 9, no. 10, pp. 1153–1168, Oct. 2002.

[10] D. Bartz, D. Mayer, J. Fischer, S. Ley, A. del Rio, S. Thust, C. Heussel, H. Kauczor, and W. Straßer, "Hybrid segmentation and exploration of the human lungs," in *Visualization, 2003. VIS 2003. IEEE*, 19-24 Oct. 2003, pp. 177–184.

[11] J. Tschirren, E. Hoffman, G. McLennan, and M. Sonka, "Intrathoracic airway trees: segmentation and airway morphology analysis from low-dose CT scans," *IEEE Transactions on Medical Imaging*, vol. 24, no. 12, pp. 1529–1539, Dec. 2005.

[12] B. van Ginneken, W. Baggerman, and E. van Rikxoort, "Robust segmentation and anatomical labeling of the airway tree from thoracic CT scans," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 5241, 2008, pp. 219–226.

[13] A. Fabijańska, "Two-pass region growing algorithm for segmenting airway tree from MDCT chest scans," *Computerized Medical Imaging and Graphics*, vol. 33, no. 7, pp. 537 – 546, 2009.

[14] P. Lo, J. Sporring, and M. de Bruijne, "Multiscale vessel-guided airway tree segmentation," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 323–332.

[15] P. Lo, J. Sporring, H. Ashraf, J. J. H. Pedersen, and M. de Bruijne, "Vessel-guided airway tree segmentation: A voxel classifcation approach." *Medical Image Analysis*, vol. 14, no. 4, pp. 527–538, Aug 2010.

[16] J. Pu, C. Fuhrman, W. F. Good, F. C. Sciurba, and D. Gur, "A differential geometric approach to automated segmentation of human airway tree," *IEEE Transactions on Medical Imaging*, vol. 30, no. 2, pp. 266–278, 2011.

[17] C. Fetita, F. Preteux, C. Beigelman-Aubry, and P. Grenier, "Pulmonary airways: 3-D reconstruction from multislice CT and clinical investigation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 11, pp. 1353–1364, 2004.

[18] P. Lo, J. Sporring, J. J. H. Pedersen, and M. de Bruijne, "Airway tree extraction with locally optimal paths," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 5762, 2009, pp. 51–58.

[19] M. Ceresa, X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solorzano, "Automatic leakage detection and recovery for airway tree extraction in chest CT images," in *Proc. IEEE Int Biomedical Imaging: From Nano to Macro Symp*, 2010, pp. 568–571.

[20] G. Song, N. Tustison, and J. C. Gee, "Airway tree segmentation by removing paths of leakage," in *Proc. of Third International Workshop on Pulmonary Image Analysis*, 2010, pp. 109–116.

[21] H. Singh, M. Crawford, J. P. Curtin, and R. Zwiggelaar, "Automated 3D segmentation of the lung airway tree using gain-based region growing approach," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 3217, 2004, pp. 975–982.

[22] K. Lai, P. Zhao, Y. Huang, J. Liu, C. Wang, H. Feng, and C. Li, "Automatic 3D segmentation of lung airway tree: A novel adaptive region growing approach," in *Proc. of the 3rd International Confeference on Bioinformatics and Biomedical Engineering*, 2009, pp. 1–4.

[23] D. Babin, E. Vansteenkiste, A. Pižurica, and W. Philips, "Segmentation of airways in lungs using projections in 3-D CT angiography images," in *Proc. of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2010, pp. 3162–3165.

[24] D. Aykac, E. Hoffman, G. McLennan, and J. Reinhardt, "Segmentation and analysis of the human airway tree from three-dimensional X-ray CT images," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 940–950, 2003.

[25] B. Irving, P. Taylor, and A. Todd-Pokropek, "3D segmentation of the airway tree using a morphology based method," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 297–307.

[26] C. Fetita, M. Ortner, P.-Y. Brillet, F. Prêteux, and P. Grenier, "A morphological-aggregative approach for 3D segmentation of pulmonary

airways from generic MSCT acquisitions," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 215–226.

[27] R. Pinho, S. Luyckx, and J. Sijbers, "Robust region growing based intrathoracic airway tree segmentation," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 261–271.

[28] M. Feuerstein, T. Kitasaka, and K. Mori, "Adaptive branch tracing and image sharpening for airway tree extraction in 3-D chest CT," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 273–284.

[29] A. Fabijańska, "Results of applying two-pass region growing algorithm for airway tree segmentation to MDCT chest scans from EXACT database," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 251–260.

[30] C. Bauer, T. Pock, H. Bischof, and R. Beichel, "Airway tree reconstruction based on tube detection," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 203–213.

[31] C. S. Mendoza, B. Acha, and C. Serrano, "Maximal contrast adaptive region growing for CT airway tree segmentation," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 285–295.

[32] R. Wiemker, T. Bülow, and C. Lorenz, "A simple centricity-based region growing algorithm for the extraction of airways," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 309–314.

[33] J. Lee and A. P. Reeves, "Segmentation of the airway tree from chest CT using local volume of interest," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 333–340.

[34] S. Born, D. Iwamaru, M. Pfeif e, and D. Bartz, "Three-step segmentation of the lower airways with advanced leakage-control," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 239–250.

[35] O. Weinheimer, T. Achenbach, and C. Düber, "Fully automated extraction of airways from CT scans based on self-adapting region growing," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 315–321.

[36] C. Bauer, H. Bischof, and R. Beichel, "Segmentation of airways based on gradient vector f ow," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 191–201.

[37] E. M. van Rikxoort, W. Baggerman, and B. van Ginneken, "Automatic segmentation of the airway tree from thoracic CT scans using a multi-threshold approach," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 341–349.

[38] J. Tschirren, T. Yavarna, and J. Reinhardt, "Airway segmentation framework for clinical environments," in *Proc. of Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 227–238.

[39] P. Lo, B. van Ginneken, J. Reinhardt, and M. de Bruijne, "Extraction of airways from CT (EXACT'09)," in *Second International Workshop on Pulmonary Image Analysis*, 2009, pp. 175–189.

[40] T. Schlathölter, C. Lorenz, I. C. Carlsen, S. Renisch, and T. Deschamps, "Simultaneous segmentation and tree reconstruction of the airways for virtual bronchoscopy," in *Medical Imaging 2002: Image Processing*, M. Sonka and J. M. Fitzpatrick, Eds., vol. 4684, no. 1. SPIE, 2002, pp. 103–113.

[41] J. N. Tsitsiklis, "Eff cient algorithms for globally optimal trajectories," *IEEE Transactions on Automatic Control*, vol. 40, no. 9, pp. 1528–1538, 1995.

[42] R. Malladi and J. Sethian, "Level set and fast marching methods in image processing and computer vision," in *Proc. International Conference on Image Processing*, vol. 1, 1996, pp. 489–492 vol.1.

[43] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector f ow," *IEEE Transaction on Image Processing*, vol. 7, no. 3, pp. 359–369, Mar. 1998.

[44] T. Heimann, B. van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. M. M. Cashman, Y. Chi, A. Córdova, B. M. Dawant, M. Fidrich, J. D. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmller, R. I. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H.-P. Meinzer, G. Németh, D. S. Raicu, A.-M. Rau, E. M. van Rikxoort, M. Rousson, L. Ruskó, K. A. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. M. Waite, A. Wimmer, and I. Wolf, "Comparison and evaluation of methods for liver segmentation from CT datasets." *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1251–1265, Aug 2009.

[45] M. Niemeijer, M. Loog, M. D. Abramoff, M. A. Viergever, M. Prokop, and B. van Ginneken, "On combining computer-aided detection systems." *IEEE Transactions on Medical Imaging*, vol. 30, no. 2, pp. 215–223, Feb 2011.