Evaluation of Registration Methods on Thoracic CT: The EMPIRE10 Challenge

Keelin Murphy, Bram van Ginneken*, *Member, IEEE,* Joseph M. Reinhardt, *Senior Member, IEEE,* Sven Kabus, Kai Ding, Xiang Deng, Kunlin Cao, Kaifang Du, Gary E. Christensen, Vincent Garcia, Tom Vercauteren, Nicholas Ayache, Olivier Commowick, Grégoire Malandain, Ben Glocker, *Member, IEEE,* Nikos Paragios, *Fellow, IEEE,*, Nassir Navab, *Member, IEEE,* Vladlena Gorbunova, Jon Sporring, Marleen de Bruijne, Xiao Han *Senior Member, IEEE*, Mattias P. Heinrich, Julia A. Schnabel, *Member, IEEE,* Mark Jenkinson, *Member, IEEE,* Cristian Lorenz, Marc Modat, Jamie R. McClelland, Sébastien Ourselin, Sascha E.A. Muenzing, Max A. Viergever, *Fellow, IEEE,* Dante De Nigris, D. Louis Collins, Tal Arbel, Marta Peroni, Rui Li, Gregory C. Sharp, Alexander Schmidt-Richberg, Jan Ehrhardt, René Werner, Dirk Smeets, Dirk Loeckx, Gang Song, Nicholas Tustison, Brian Avants, James C. Gee, Marius Staring, Stefan Klein, Berend C. Stoel, Martin Urschler, Manuel Werlberger, Jef Vandemeulebroucke, Simon Rit, David Sarrut, and Josien P.W. Pluim, *Senior Member, IEEE*

***Abstract*—EMPIRE10 (Evaluation of Methods for Pulmonary Image REgistration 2010) is a public platform for fair and meaningful comparison of registration algorithms which are applied to a database of intra-patient thoracic CT image pairs. Evaluation of non-rigid registration techniques is a non trivial task. This is compounded by the fact that researchers typically test only on their own data, which varies widely. For this reason, reliable assessment and comparison of different registration algorithms has been virtually impossible in the past. In this work we present the results of the launch phase of EMPIRE10, which comprised the comprehensive evaluation and comparison of 20 individual algorithms from leading academic and industrial research groups. All algorithms are applied to the same set of 30 thoracic CT pairs. Algorithm settings and parameters are chosen by researchers expert in the configuration of their own method and the evaluation is independent, using the same criteria for all participants. All results are published on the EMPIRE10 website (http://empire10.isi.uu.nl). The challenge remains ongoing and open to new participants. Full results from 24 algorithms have been published at the time of writing. This article details the organisation of the challenge, the data and evaluation methods and the outcome of the initial launch with 20 algorithms. The gain in knowledge and future work are discussed.**

***Index Terms*—Registration, Chest, Computed Tomography, Evaluation.**

Manuscript received , 2011; revised , 2011; accepted , 2011. First published , 2011; current version published , 2011. *Asterisk indicates corresponding author*.

Keelin Murphy, Sascha E.A. Muenzing, Max A. Viergever, and Josien P.W. Pluim are with the Image Sciences Institute, University Medical Center, Utrecht, The Netherlands.

*Bram van Ginneken is with the Department of Radiology, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands (e-mail: b.vanginneken@rad.umcn.nl). Bram van Ginneken is also with the Image Sciences Institute, University Medical Center, Utrecht, The Netherlands.

Joseph M. Reinhardt, Kai Ding, and Kaifang Du are with the Department of Biomedical Engineering, The University of Iowa, USA.

Sven Kabus and Cristian Lorenz are with Philips Research, Hamburg, Germany.

Xiang Deng is with Corporate Technology, Siemens Ltd., China.

Kunlin Cao and Gary E. Christensen are with the Department of Electrical and Computer Engineering, The University of Iowa, USA.

Vincent Garcia, Nicholas Ayache, and Grégoire Malandain are with INRIA Sophia Antipolis - Méditerranée, France.

Tom Vercauteren is with Mauna Kea Technologies, Paris, France.

Olivier Commowick is with INRIA Rennes - Bretagne Atlantique, France.

Ben Glocker and Nassir Navab are with Computer Aided Medical Procedures (CAMP), TU München, Germany.

Nikos Paragios is with Laboratoire MAS, Ecole Centrale Paris, Chatenay-Malabry, France, and with Equipe GALEN, INRIA Saclay - Ile-de-France, Orsay, France.

Marta Peroni, Rui Li, and Gregory C. Sharp are with Massachusetts General Hospital, Boston, USA. Marta Peroni is also with Politecnico di Milano, Milan, Italy, and with Massachusetts Institute of Technology, Cambridge, USA. She is supported by the Roberto Rocca Foundation.

Vladlena Gorbunova, Jon Sporring, and Marleen de Bruijne are with Department of Computer Science, University of Copenhagen, Denmark. Marleen de Bruijne and Stefan Klein are with the Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, The Netherlands.

Xiao Han is with CMS Software, Elekta Inc., USA.

Mattias P. Heinrich and Julia A. Schnabel are with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK. Mattias P. Heinrich and Mark Jenkinson are with the Oxford University Centre for Functional MRI of the Brain, Oxford, UK.

Marc Modat, Jamie R. McClelland, and Sébastien Ourselin are with the Centre for Medical Image Computing, University College London, London, UK.

Dante De Nigris and Tal Arbel are with the Centre for Intelligent Machines, McGill University, Canada.

D. Louis Collins is with the Department of Biomedical Engineering, McGill University, Canada.

Alexander Schmidt-Richberg, Jan Ehrhardt, and René Werner are with the Institute of Medical Informatics, University of Lübeck, Germany.

Dirk Smeets and Dirk Loeckx are with K.U. Leuven, Faculty of Engineering, ESAT/PSI, Medical Imaging Research Center, Belgium.

Gang Song, Brian Avants, and James C. Gee are with Penn Image Computing and Science Laboratory (PICSL), Department of Radiology, University of Pennsylvania School of Medicine, USA.

Nicholas Tustison is with the Department of Radiology and Medical Imaging, University of Virginia, USA.

Marius Staring and Berend C. Stoel are with the Division of Image Processing, Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands.

Martin Urschler and Manuel Werlberger are with the Institute for Computer Graphics and Vision, Graz University of Technology, Austria. Martin Urschler is also with Ludwig-Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria.

Jef Vandemeulebroucke, Simon Rit, and David Sarrut are with the Université de Lyon, CREATIS, CNRS UMR5220, Inserm U1044, INSA-Lyon, Université Lyon 1, Centre Léon Bérard, France.

I. Introduction

FOR many years researchers have worked on registration algorithms for medical imaging applications [1], [2], [3], [4], [5], [6]. One such application is the alignment of thoracic CT images from the same subject, in particular of the lung and its internal structures. The lungs are highly deformable organs making accurate registration of them a challenging task requiring a non-rigid registration approach [7]. There are many scenarios in which intra-patient pulmonary registration is clinically useful. Registration of follow-up (temporally distinct) breathhold inspiration scans should make visual comparison of these scans a much easier and less error-prone task for a radiologist. For well-aligned images, automatic methods of comparison for analysis of disease progression etc. may even be considered. Breathhold inspiration scans may also be aligned and compared with breathhold expiration scans to enable improved monitoring of airflow and pulmonary function via CT images. Where 4D data is available (i.e. numerous CT images representing various phases in a breathing cycle) these images may be registered in order to obtain information about the deformations that occur during respiration. Such information can be used in image-guided treatment, including motion estimation in treatment planning and is also expected to be extremely valuable in understanding the effects of disease on (regional) lung elasticity.

The inability to compare registration algorithms in a meaningful way is a major obstacle to further development and improvement in the research community. Although many researchers have published articles demonstrating the results of their registration algorithms, they are largely based on proprietary datasets, even with differing image modalities. Furthermore their methods of evaluating their registrations, which is a highly complex task in itself, are diverse, further complicating the task of comparing algorithm results. Some authors have undertaken the task of running a number of different algorithms on a fixed dataset in order to compare the algorithm performances in a reliable manner [8], [9], [10], [11], [12], [13]. The drawback to this approach, however, is that the configuration of algorithm parameters for a specific task is frequently a non-trivial problem which is best understood by those who developed the method. Ideally the algorithm should be implemented and configured by those who are thoroughly familiar with all aspects of its behaviour in order to obtain optimal performance. There have been some initiatives in the past which provided common datasets and evaluation methods for the evaluation of registrations of brain images [14], [15], [16], while allowing the users to configure and run their own registration algorithms on the data. An attempt was made to provide an objective comparison of pulmonary registration algorithms in [17] and [18] but based on just one pair of lung images in the case of [17] and a single phantom in the case of [18]. Furthermore the methods of registration evaluation in those works are limited to analysis of 38 manually identified landmarks [17] and 48 plastic markers [18]. Results from 12 algorithms are reported in [17] and from 8 algorithms in [18].

The EMPIRE10 (Evaluation of Methods for Pulmonary Image REgistration 2010) challenge [19] described in this article provides a means for objective comparison of registration algorithms applied to 30 pairs of thoracic CT data. This challenge invites participants to download a set of thoracic CT intra-patient scan pairs and register them using their own registration algorithms. The aim of the registration is to align the lung volumes; structures outside the lungs are not considered during the registration evaluation. The scans have been selected by the organisers to represent a broad variety of problems of the type encountered in clinical practice. Participants calculate deformation fields and submit them to the EMPIRE10 organisational team for independent evaluation. The deformation fields are evaluated over four individual categories: Lung boundary alignment, fissure alignment, correspondence of manually annotated point pairs and the presence of singularities in the deformation field. Evaluation results are published on the EMPIRE10 website [19]. The advantages of this approach to registration evaluation are as follows:

- All algorithms will be applied to exactly the same set of data, designed to be as large and diverse as possible.
- Any algorithm parameters or settings will be chosen by those familiar with the algorithm and expert in its configuration.
- The resulting registrations will be independently evaluated, in 4 different categories, using the same criteria for all participants.

This article describes the organisation of the challenge and its initial two-phase launch. Phase 1 required participants to register 20 data sets in their own facilities and return their registration results to the challenge organisers for evaluation. Phase 2 consisted of a live workshop at the MICCAI conference in 2010 [20], where participants registered a further 10 scan pairs. The aim of this work is to describe the challenge in detail and discuss the outcome of phases 1 and 2 and the advancement in knowledge achieved.

II. Materials

The materials for this challenge were gathered from several sources to try to include as broad a variety as possible of the scenarios encountered in clinical practice. Thus, scans may be taken at various phases in the breathing cycle (full breath-hold inspiration, full breath-hold expiration, phase from 4D breathing data). Subjects may exhibit lung disease or appear healthy, although they typically do not exhibit gross pathology. Data from a variety of scanners is included and a variety of different slice-spacings occur.

In this section we describe in detail the properties of the 30 scan pairs provided to participants. Each scan pair is taken from a single subject, i.e. only intra-patient registration is considered in this challenge. The lungs in all images were segmented using an automatic algorithm from van Rikxoort et al. [21]. Lung segmentations were checked and altered manually where necessary. In all cases the scan data was cropped using a bounding box around the lungs before distribution. This was done to reduce the size of the files to be downloaded since the regions outside the lungs were to be excluded from consideration during registration and evaluation. The data downloaded by participants also included the binary lung masks which they were permitted to use during registration. No other segmentation information was provided.

The remainder of this section describes the 30 scan pairs categorised by type. Table I lists which scan pairs belong to which category. Note that the participants were not aware which scans belonged to which category, or even what categories of data were included, until after they had registered the scans and their results had been published.

A. Breathhold Inspiration Scan Pairs

Eight of the thirty scan pairs consisted of two breathhold inspiration scans (referred to as 'insp-insp' in table I). These scans were acquired as part of the Nelson Study [22] which is the largest lung cancer screening trial in Europe. Current and former heavy smokers, mainly men, aged 50 to 75 years are included in this study. In these 8 pairs the follow up scans were made between 9 and 14 months after the baseline scan. A low-dose protocol was used (30mAs) and the scanner was either Philips Brilliance 16P or Philips Mx8000 IDT 16 in each case. Slice thickness was 1.00 mm with slice-spacing of 0.70 mm. Pixel spacing in the X and Y directions varied from 0.68 mm to 0.78 mm with an average of 0.74 mm.

B. Breathhold Inspiration and Expiration Scan Pairs

A further 8 scan pairs, also taken from the Nelson Study [22] were made up of a breathhold inspiration scan and a breathhold expiration scan, made in the same session (referred to as 'insp-exp' in table I). The inspiration scan was created using a low-dose protocol (30mAs) while the expiration scan was ultra-low-dose (20mAs). The scanner used was Philips Brilliance 16P with slice thickness of 1.00 mm and slice spacing of 0.70 mm. Pixel spacing in the X and Y directions varied from 0.63 mm to 0.77 mm with an average value of 0.70 mm.

C. 4D Data Scan Pairs

Four of the scan pairs consisted of two individual phases from a 4D dataset. In each case the phases were chosen to be as distinct as possible, i.e. at opposing ends of the breathing cycle. Three of the scan pairs were from a GE Discovery ST multislice PET/CT scanner while the fourth (scan pair *17*)[23] was from a Philips Brilliance CT 16 Slice scanner. The scans from the GE scanner used a beam current of 100mAs each, while the Philips scan used 400mAs. Since each scan pair came from a 4D dataset the spacing was identical for the two scans in the pair. Slice-spacing was 1.25 mm, 2.50 mm and 2.50 mm for the 3 GE scans and 2.00 mm for the Philips scan. Pixel spacing in the X and Y directions was set at 0.98 mm in all cases.

D. Ovine Data Scan Pairs

Four scan pairs were ovine (sheep) data from two datasets where breathing was regulated. A number of metallic markers (67 in the first animal, 103 in the second), 1.40 mm in diameter, had been surgically implanted in the sheep lungs approximately 6 weeks before scanning. The markers were implanted mainly in the left upper lobe and right lower lobe. Airway pressure was regulated during scanning on a Philips MX8000 Quad Scanner with the sheep in supine position. Scans were acquired at 3 different airway pressures: 8, 16 and 24 cm H_2O. Slice spacing was 0.60 mm with in-plane pixel spacing of 0.47 mm the first animal and 0.49 mm for the second.

The metallic markers which were visible in the scans were identified and their locations noted. They were subsequently disguised using a hole filling technique in order that participants could not identify them and registration algorithms would not be guided by them. The marker locations were used in the registration evaluations (see section IV-C) .

E. Contrast - Non-Contrast Scan Pairs

Two pairs of scans were used in which contrast material was present in one scan of the pair but not in the other. These scans were acquired on a Siemens SOMATOM Sensation CT 64-slice scanner. The contrast scan (arterial phase) was acquired approximately 30 seconds after the non-contrast scan in each case. Slice spacing was 1.50 mm with pixel spacing in the X and Y directions of 0.60 mm for the first subject and 0.69 mm for the second.

F. Artificially Warped Scan Pairs

Since registration algorithms are difficult to evaluate in a quantitative way, a frequently employed method (e.g. [24], [25], [26]) is to apply a known artificial transformation to a single dataset and then attempt to register the original scan with the result. In this case the ground truth is known so evaluation is more reliable. For this reason 4 scan pairs were included in the EMPIRE10 challenge which consisted of an original scan and the same scan with an artificial thin-plate-spline warp applied to it.

The procedure for warping a scan artificially was as follows: A pair of breathhold inspiration scans from the Nelson Study [22] was acquired. One hundred well-dispersed landmark points were identified automatically in the baseline scan and matched semi-automatically in the follow-up scan. Landmark identification and matching was done according to the method described in [27], [28]. A thin-plate-spline model was created using the 100 pairs of matching points. Using this thin-plate-spline model and linear interpolation, the baseline image was warped to create an image with the same image size and spacing as the follow up scan. The anatomical appearance of this warped scan was, by construction, similar to that of the follow up scan. This method was used in order to ensure that the artificial warp would result in an image with a realistic appearance. A sharpening filter (unsharp masking) was applied to the warped image to negate the smoothing effects of warping and interpolation. Regions around the edge of the warped image (outside the lungs) where no data values could be assigned were cropped away. The scan pair distributed to the challenge participants in each case consisted of the original baseline scan and the artificially warped version of this scan.

Pair	Data Category	Pair	Data Category	Pair	Data Category
01	Insp-Exp	11	Insp-Insp	21	Insp-Exp
02	Insp-Insp	12	Warped	22	Insp-Insp
03	Insp-Insp	13	4D	23	4D
04	Ovine	14	Insp-Exp	24	Ovine
05	Warped	15	Insp-Insp	25	Warped
06	Contrast	16	4D	26	Contrast
07	Insp-Exp	17	4D	27	Insp-Insp
08	Insp-Exp	18	Insp-Exp	28	Insp-Exp
09	Insp-Insp	19	Insp-Insp	29	Ovine
10	Ovine	20	Insp-Exp	30	Warped

TABLE I
A LISTING OF WHICH CATEGORY OF DATA WAS PROVIDED FOR EACH OF SCAN PAIRS 01 TO 30. EXPLANATIONS OF THE DATA CATEGORIES ARE GIVEN IN SECTION II.

The scans were acquired using either Philips Mx8000 IDT 16 or Philips Brilliance 16P scanners. Slice-spacing was 0.70 mm while in-plane pixel spacing varied from 0.66 mm to 0.80 mm with an average value of 0.74 mm.

III. CHALLENGE SETUP

The EMPIRE10 challenge was launched in April 2010 when a large number of researchers involved in the fields of registration and thoracic CT (as determined by a literature search) were invited by email to visit the website [19] and to participate in the challenge. The challenge was also widely announced on mailing lists. The registration tasks involved were divided into two phases described below. The two phases are considered independently in this work since the circumstances of registering were generally different for each. The participants were not given any information about the source or type of data they were registering until after they had completed the registrations and submitted their results.

- Phase 1: The participants downloaded 20 pairs of thoracic CT scans (pairs *01-20* in table I) from the 30 pairs described in section II. These pairs were registered by the participants in their own facilities, and results in the form of deformation field images were submitted to the organisers by June 14th. These registrations were evaluated (see section IV) and the results were published on the website [19].
- Phase 2: The participants took part in the Grand Challenge Workshop [20] at the MICCAI [29] conference in Beijing on September 24th 2010. During the first 3 hours of the workshop participants were required to register the remaining 10 datasets (pairs *21-30* in table I) which had been password encrypted until that point. Since registration of such large datasets is technically challenging in terms of processing power and memory requirements it was permitted to perform the registrations using remotely located hardware if required. If a participant was unable to attend, or unable to complete registration by the end of the three hours it was permitted to submit results during the week following. (Note that it was not permitted to submit partially complete results at the workshop and supplement these with additional results during the week following). Algorithms which were run remotely or whose results were not submitted during the workshop but rather in the week following are clearly noted in the results section as well as on the challenge website.

Since September 2010 the EMPIRE10 challenge has entered a new ongoing phase and remains open to entries from new participants or to submission of improved results from teams already involved. In this way we hope that EMPIRE10 will continue to maintain a record of the state of the art in registration of thoracic CT. All results published on the website now are based on the combined set of 30 scan pairs. The individual sets of results from phase 1 and phase 2 as described in this article remain on the website but are reported separately for reference only. Latest results on the combined 30 datasets, some of which have been recently updated after algorithm modifications, can be found on the EMPIRE10 website [19].

IV. EVALUATION

Evaluation of registration algorithms was carried out in four different ways as described in the remainder of this section. Note that for the EMPIRE10 challenge the image to be deformed is referred to as the 'moving image' while the reference image is known as the 'fixed image'.

Participants were asked to declare whether their method was fully automatic (processed all scan pairs with the same parameter set), semi-automatic (required different parameters for different scan pairs), or interactive (required more significant user interaction such as manual alignment, defining corresponding point pairs etc.) and this information is shown on the challenge website [19] as well as in the results section.

A. Alignment of Lung Boundaries

Aligning the boundaries of the lungs correctly is one of the most fundamental expectations of a pulmonary CT registration algorithm. The lung boundary is easily defined in CT in most regions, with the notable exception of the mediastinal (central) region. A method of analysing lung boundary registration was therefore developed for this challenge, which is restricted to the peripheral regions where the obvious density change between lung parenchyma and chest wall occurs.

The lungs in all images were segmented using an automatic algorithm from van Rikxoort et al. [21]. Lung segmentations were checked and altered manually where necessary. All further processing described in this section was carried out fully automatically. The lung boundary defined by the lung segmentations was extracted and a distance transform image was generated from the boundary image. The mediastinal region of the lung boundaries was masked out as follows (see figure 1): The centre-of-mass of both lungs combined, $cMass$ was determined. The Euclidean distances from $cMass$ to the centre-of-mass of the left lung, and to the centre of mass of the right lung were determined. A sphere centred at $cMass$ and with a radius defined by the larger of these two distances was used to identify locations close to the mediastinum. All voxels within this sphere were excluded from further processing.

Next, points within 20 mm of the lung boundary were marked in order to define boundary adjacent locations. Points within 2 mm of the boundary were excluded to allow for

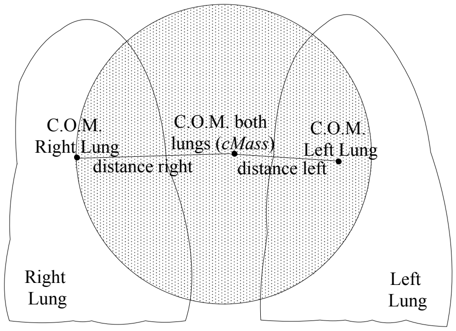


Fig. 1. A schematic representation of the method to mask out the mediastinal region as described in section IV-A. The abbreviation C.O.M. refers to the centre of mass. The shaded region in the diagram is excluded from lung boundary evaluation due to its proximity to the mediastinum. It is defined by the sphere centred at $cMass$ and with radius of either 'distance right' or 'distance left' - whichever is larger.

minor inaccuracies in the lung segmentation. [1] Points inside and outside the lung boundaries were distinguished using the lung segmentation image and marked with different values v_{in} and v_{out} respectively. These markings constituted the reference standard for checking lung boundary alignment. See figure 2(a) as an example.

Each participant submitted deformation field data for each registration carried out. Using this data, it was calculated for each point p_{fixed} marked with v_{in} or v_{out} in the fixed image, which point p_{reg} in the moving image was aligned with this location. If p_{fixed} was marked with v_{in} and p_{reg} was marked with v_{out} then a unit penalty was incurred. Similarly the reverse situation where p_{fixed} was marked with v_{out} and p_{reg} was marked with v_{in} also incurred a unit penalty. Note that if p_{reg} was not marked with either v_{in} or v_{out} (i.e. if it lay within 2 mm of the boundary, or more than 20mm from the boundary) then no action was taken.

Error in lung boundary alignment was calculated as the percentage of points marked with v_{in} or v_{out} which were registered to points marked as being on the opposite side of the boundary. This value was given as the overall score in the lung boundary alignment category. For information, the errors in the left lung, right lung, upper lung and lower lung were also calculated and displayed on the participant's results page on the challenge website [19].

B. Alignment of Major Fissures

Fissures are plate-like structures which divide the lungs into regions called lobes. Since fissures represent important physical boundaries within the lungs their alignment is included as an evaluation category in the EMPIRE10 challenge. To simplify the evaluation, particularly for poor quality data where minor fissure structures may be difficult to see, we evaluate the registration of the major fissures only. Each lung contains a single major fissure dividing it into an upper and a lower section. This method of analysis was developed specifically for use in the EMPIRE10 challenge.

The fissures in all images were segmented using an automatic algorithm from van Rikxoort et al. [30]. Fissure segmentations were checked and altered manually to exclude minor fissures and any erroneous markings. Gaps in the segmentation were not always filled so the resulting segmentation may be incomplete but does not contain any non-fissure structures. All further processing described in this section was carried out fully automatically.

A distance transform image was generated from the resulting fissure segmentation image. Next, points within 20 mm of the fissure segmentation were marked, excluding those within 2 mm of the fissure to allow for minor inaccuracies in the segmentation. Points which were not directly above or below a fissure voxel (looking in the axial direction) were excluded in order to prevent the marked regions wrapping around the edges of the fissure plates (or around gaps in incomplete fissure segmentations). For each marked point p, the closest point p_{fiss} on the fissure segmentation was determined. Points above and below the fissure are distinguished by comparing the axial components of p and p_{fiss}. Different values, v_{above} and v_{below} were used to mark points above and below the fissure respectively. These markings constituted the reference standard for checking fissural alignment. See figure 2(b) as an example.

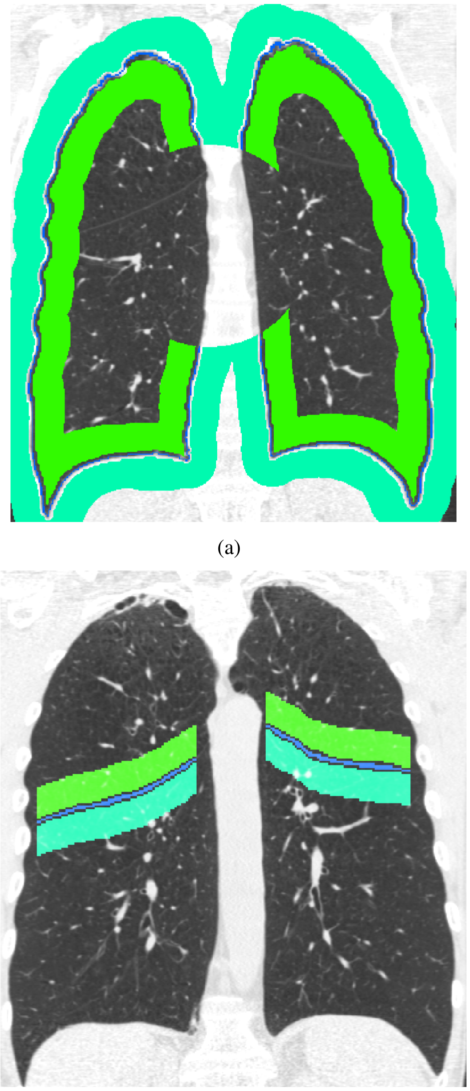
Using the deformation data submitted by the participant, it was calculated for each point p_{fixed} marked with v_{above} or v_{below} in the fixed image which point p_{reg} in the moving image was aligned with this location. If p_{fixed} was marked with v_{above} and p_{reg} was marked with v_{below} then a unit penalty was incurred. Similarly the reverse situation where p_{fixed} was marked with v_{below} and p_{reg} was marked with v_{above} also incurred a unit penalty. Note that if p_{reg} was not marked with either v_{in} or v_{out} (i.e. if it lay within 2 mm of the boundary, or more than 20mm from the boundary) then no action was taken.

Error in fissure alignment was calculated as the percentage of points marked with v_{in} or v_{out} which were registered to points marked as being on the opposite side of the boundary. This value was given as the overall score in the fissure alignment category. For information, the errors in the left lung and right lung were also calculated and displayed on the participant's results page on the challenge website [19].

C. Correspondence of Annotated Landmark Pairs

Analysis of point correspondence is a commonly used way to evaluate registration algorithms [31], [13], [32], [33], [34], [35], [36]. In most cases the set of points is manually annotated by an expert, resulting in relatively small point sets, which are frequently clustered in the mediastinal region where distinctive anatomical points are more easily observed. For the EMPIRE10 challenge a well-distributed set of 100 distinctive landmark points was automatically defined in the fixed image from each scan pair. Each point p_{fixed} was then matched with

[1] This 2 mm margin, mentioned in both sections IV-A and IV-B is chosen as it is assumed that any segmentation error larger than 2mm could not have been overlooked during the segmentation checking process. Making the margin smaller would certainly detect more errors in the registration results, many of which would be legitimate. However it would also risk penalising some algorithms unfairly where the error lay with the segmentation and not with the registration.



(a)

(b)

Fig. 2. **(a)** Coronal section of the lung boundary reference standard. The boundary itself is marked in blue and surrounded by a 2 mm gap on each side. Regions outside the lung are marked in cyan, and inside the lung are marked in green. **(b)** Fissure reference standard. Colour coding is analogous to that in the left hand image with regions above and below the major fissures marked in green and cyan respectively.

the corresponding point p_{moving} in the moving image using a semi-automatic method. The methods for defining and matching the points are described in Murphy et al. [27], [28]. The software used is publicly available at http://isimatch.isi.uu.nl. An example of the point distribution is shown in figure 3(a). The landmarks are designed to be well dispersed throughout the lungs and, in most cases, lie in regions of good contrast (to enable them to be visually matched), typically on the boundary of vessel and parenchyma. The manual component of the point-matching procedure allows the user to determine the match by examination of the point in all three orthogonal directions and at various zoom levels. A matching point may be selected or moved at any time. Corresponding points were marked by either 3 or 4 observers independently (from 7 available observers who worked on this task), and any location where any pair of observer opinions differed by 3 mm or more was checked a final time by an observer who could see all previous annotations on a single screen and accept or reject each one independently. The observers were all medical students, except for one radiologist in training. All observers received instruction, training and practice in this task before beginning.) The rejected points were not included in the reference standard, all other points were retained. (If all annotations for a landmark were rejected the landmark itself was excluded. For this reason 7 scan pairs were left with only 99 annotated landmarks and 1 scan pair with 98.) An example of a landmark with several observer opinions is shown in figure 4. By accepting more than one observer opinion as truth, we acknowledge that in most cases it is not possible to identify a matching point with perfect accuracy. This is related to many issues such as image quality, voxel size and the partial volume effect.

The deformation data submitted by each participant was used to calculate for each of the defined points p_{fixed} in the fixed image which point p_{reg} in the moving image was aligned with this location. The point p_{reg} was then compared (using Euclidean distance) with the reference standard point p_{moving}. Where several acceptable options for p_{moving} were defined, the p_{moving} that was closest to p_{reg} was used as the reference. Note that p_{reg} was rounded to the nearest voxel location before distance calculation. Since all observer marks were made without sub-voxel accuracy, an algorithm which agrees precisely with a particular observer may therefore obtain an error of zero.

The distance d from p_{moving} to p_{reg} was calculated in mm for each of the annotated point pairs. The overall error in the landmarks category was given by the average of all the distances d in the scan-pair. For information, the minimum distance, the maximum distance, the average distances in the upper and lower lungs and the average distance in each of the three orthogonal directions (Anterior-Posterior, Superior-Inferior and Left-Right) were also calculated and displayed on the participant's results page on the website [19].

There are a number of scan pairs that were treated as special cases in terms of the evaluation using landmark pairs. For the ovine data the landmark locations were given by the fiducial markers as described in section II-D and not manually annotated as for the other data. (Therefore scan pairs 4 and 10 have 67 landmarks each while scan pairs 24 and 29 have 103). The fiducial markers do not necessarily lie on high contrast boundaries, see figure 5 as an example.

Furthermore, for the artificially warped data (see section II-F) the landmark pairs which were used to specify the thin-plate-spline model were used as the reference standard in landmark evaluation, meaning that just one (completely precise) matching point was available for each landmark defined.

To demonstrate the level of accuracy of the points which were annotated using the semi-automatic system described in [27], [28] the mean and standard deviation of the inter-observer distances for each of the 22 scan pairs concerned are provided in table II.

Pair	01	02	03	06	07	08	09	11	13	14	15	16	17	18	19	20	21	22	23	26	27	28
Mean	1.4	0.2	0.2	0.3	1.2	0.5	0.2	0.3	0.4	1.1	0.2	0.5	0.4	0.9	0.2	1.4	2.2	0.4	0.5	0.2	0.2	1.2
StdDev	1.8	0.5	0.5	0.6	1.4	1.0	0.7	0.7	0.7	1.6	0.5	0.8	0.7	1.3	0.6	1.4	2.3	0.9	0.8	0.5	0.6	1.3

TABLE II

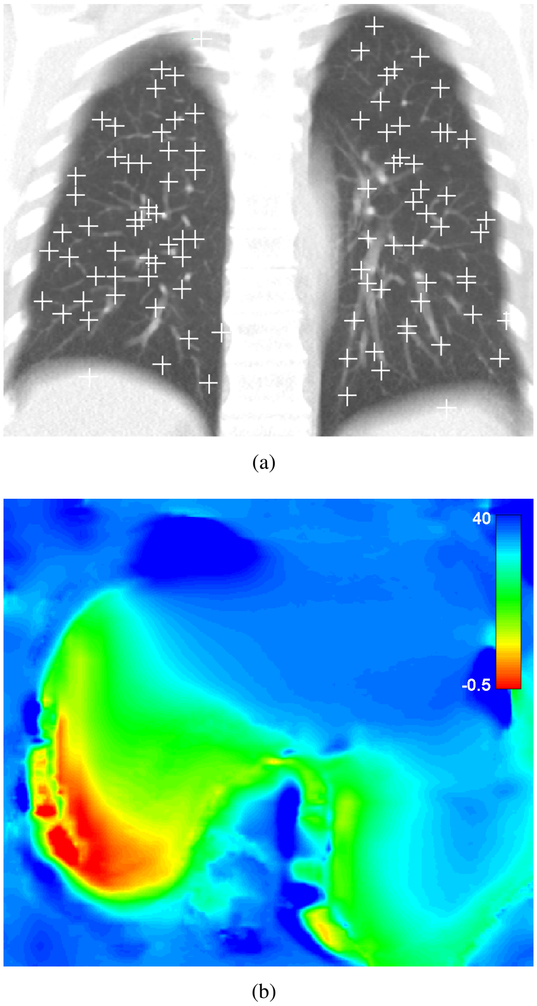
STATISTICS RELATING TO INTER-OBSERVER DISTANCES FOR MATCHING POINT PAIRS WHICH WERE DEFINED USING THE SEMI-AUTOMATIC SYSTEM DESCRIBED IN [27], [28]. ALL MEASUREMENTS ARE IN MM. THE TOP ROW SHOWS THE PAIR ID, THE SECOND ROW SHOWS THE MEAN OF THE INTER-OBSERVER DISTANCES AND THE THIRD ROW SHOWS THE STANDARD DEVIATIONS.

D. *Singularities in the Deformation Field*

The final category of evaluation is designed to analyse how physically plausible the registration deformation is. Some registration algorithms may appear to align visible structures very well, but in doing so may require physically impossible deformations. In particular we expect that a deformation should be bijective, i.e. define a one-to-one correspondence between points in the fixed image and points in the moving image. Regions where the deformation field is not bijective are commonly referred to as singularities.

Each participant submitted deformation field data for each registration carried out. The determinant of the Jacobian of the deformation field, j, (described well in [37]) was calculated at every point. This specified for each point whether local expansion or contraction had taken place. Where $j < 1$ local contraction is implied, $j = 1$ implies no change and $j > 1$ implies local expansion. Figure 3(b) shows an example of a colour-coded Jacobian image. All points within the lung volume were checked and any location where $j \leq 0$ was a singularity in the deformation field. For each such point a unit penalty was incurred. Points outside the lung volume were disregarded.

The overall error in the singularities category was given by the percentage of checked points for which penalties were incurred. For information, the errors in the left lung, the right lung, the upper lung and the lower lung were also calculated and reported on the participant's results page on the website [19].



(a)

(b)

Fig. 3. **a:** An example of the landmark points identified in a fixed scan. Landmarks have been projected onto a single slice (maximum intensity projection image is shown here) and markers are increased in size for visualization. **b:** A colour coded Jacobian image with the scale going from -0.5 (red) to 40 (blue). Pixels at or below 0 are singularities.

V. SCORING AND RANKING

It should be noted that although every attempt was made to evaluate algorithm performance as accurately as possible, there is nonetheless some room for minor errors in evaluation. For example, very small lung boundary alignment errors will be overlooked due to the 2mm region on each side of the lung boundary which we exclude from our evaluation in order to compensate for any minor lung segmentation errors. In addition, the corresponding point pairs identified as part of the reference standard cannot be guaranteed to be completely accurate - indeed in most cases it is not possible to match points completely accurately due to the partial volume effect. However, the scoring system has been designed to be as fair as possible and we consider that it is reasonable to rank teams based on these scores. The final rankings are calculated as follows:

Error scores in the four individual categories are calculated as described in section IV. A score is awarded to each participant for each scan-pair in each category (note that lower scores always imply better registration). Since these scores are based on independent measurements of different concepts there is no obvious way to combine them into a single participant score. A ranking system was therefore devised in order to measure a participant's overall performance and to compare participants with each other.

The ranking scheme works as follows for a theoretical group of n participants: The error score of a participant for scan-pair s and evaluation category c is compared with the corresponding error score of all other participants. The participant is then awarded a ranking r_{sc} for that scan-pair

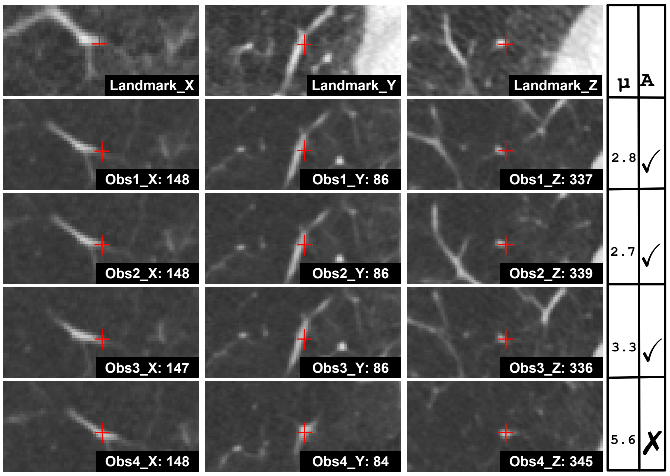


Fig. 4. An example of differing observer opinions for a landmark in scan pair 07. The top row shows the landmark identified in the fixed image in the sagittal (X), coronal (Y) and axial (Z) directions. Subsequent rows show the points selected by 4 different observers in the moving image. The slice number is shown with each orthogonal direction. The value μ shown to the right of each observer opinion is the average distance (in mm) of that point from other observer choices. The mark at A implies whether the point was accepted by the final 'checking' observer who could see all chosen points. The point chosen by observer 4 was not accepted and therefore does not form part of the reference standard.

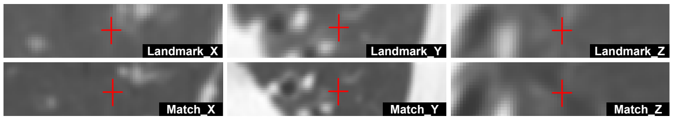


Fig. 5. A typical example of a landmark in ovine data (scan pair 04). The top row shows the landmark in the fixed scan in the sagittal (X), coronal (Y) and axial (Z) directions. The bottom row shows the matching location in the moving scan. Since these landmarks are based on the locations of fiducial markers (which have been disguised to hide them in the final images) they are not necessarily found on high contrast boundaries.

and category. Where all participants have different error scores, the participant with the lowest error will be ranked 1 while the participant with the highest error will be ranked n. If there are ties in some participant scores then the ranks must be re-arranged such that those participants rank equally. This is done as follows: Participants with equal scores initially obtain (randomly assigned) adjacent rankings. Each group of participants with equal scores is then examined, their ranks are averaged, and the average rank is assigned to each one of them. For example, scores of 0.1, 0.5, 0.5, 2 would result in rankings of 1, 2.5, 2.5, 4.

When all ranks r_{sc} have been assigned for individual scan pairs and categories they are averaged over all scan pairs to give each participant an average ranking r_c per evaluation category. (Note that because the average ranking, r_c, is based on the individual rankings, r_{sc}, and not on the average scores, a linear relationship between the average score and the average ranking is not to be expected.) Finally the per-category rankings can be averaged over the 4 evaluation categories to give the participant a final average ranking r. These final rankings are used to place the participants, with the lowest ranking in 1^{st} place and the highest in n^{th} place. If there is a tie in final rankings the placement value will be calculated by averaging in the same way as described above.

VI. Challenge Entries

Phase 1 of the challenge attracted interest from 23 teams with a combined total of 34 competing algorithms. A team was permitted to submit more than one algorithm provided that there was a significant difference between the methods, beyond a simple alteration of parameters for example.

For phase 2, 6 of these teams (with a total of 9 algorithms) declined to participate further due to other commitments. The 17 remaining teams (25 algorithms) were able to participate in phase 2, however due to restrictions on time for processing during the MICCAI Grand Challenge workshop a number of teams which had previously entered more than one algorithm decided to use only their best performing algorithm in phase 2. Ultimately, a total of 20 algorithms from the 17 teams competed in the second phase.

The remainder of this article deals only with those algorithms which were entered in both phase 1 and phase 2. Please note that rankings provided in this work for the phase 1 stage are from a total of 34 participating algorithms, although only 20 of those are being discussed here.

Below is a brief description of each of the 20 algorithms. The displayed labels A-T will be used to refer to the algorithms hereafter. For reference the corresponding algorithm name which is used on the website [19] is given here in brackets after each label. Appendix A provides explanations for commonly used registration related acronyms and abbreviations. A summary of important information for each algorithm is given in table III. For a detailed description of a particular algorithm please refer to the appropriate cited article from the proceedings of the MICCAI Grand Challenge Workshop [20].

- A (Asclepios1) [38], [39]: An initial block-matching based affine registration is applied prior to performing a diffeomorphic demons non-rigid registration. Both steps use lung masks and work in a multi-resolution manner. This method ensures that the final transformation is one-to-one.
- B (Asclepios2) [40], [39]: An affine registration followed by a non-rigid registration are applied to the scans. Both methods use lung masks and are based on a pyramidal block-matching approach. The non-rigid method is coupled with an outlier rejection procedure to improve the accuracy of the motion estimation.
- C (CMS) [41]: An affine registration is first computed followed by automatic feature detection and matching. The matched features are used to guide an MI-based block-matching image registration, the result of which is further refined by a hybrid MI/NSSD dense deformable registration procedure.
- D (DIKU) [42], [43]: A tissue appearance model based on the principle of preservation of total lung mass throughout the breathing cycle is proposed. An affine transform using extracted anatomical information is followed by a series of B-Spline transforms using mass preserving SSD as a similarity measure.

- E (DROP) [44], [45]: After initial pre-alignment, the dense intensity-based registration is performed using hierarchical FFDs and iterative discrete labeling of MRFs for the energy minimization. The energy function consists of the SAD and a first-order smoothness term.
- F (elastix) [46], [47]: A three stage approach is used: an affine step without masks followed by two non-rigid stages (B-splines, without and with masks respectively). The registration is driven by a normalized correlation metric, and optimized by a parameter free stochastic gradient descent routine.
- G (ICG LBI Graz Anisotropic Optical Flow) [48], [49]: An initial rigid registration is performed using the provided lung masks. Next a multi-scale optical flow model, consisting of SAD data and a robust Huber-norm based regularisation term, is solved using a primal-dual optimization algorithm.
- H (IMI Lübeck Diffeomorph) [50], [51]: First a nonlinear surface registration of the lungs is performed. Subsequently, an intensity-based diffeomorphic registration of the CT data is applied, using demons-like forces and diffusion regularisation. Diffeomorphisms are parameterized by static velocity fields.
- I (Iowa sstvd ssvmd Laplacian) [52], [53]: A non-rigid registration algorithm is used to match lung CT images by preserving both parenchymal tissue volume and vesselness measures in the regions of interest defined by the lung masks. The transformations are represented by B-splines and regularised using a Laplacian constraint.
- J (ISI@UMCU) [54]: A knowledge model is used to incorporate statistical information from a landmark reference set and information obtained by extracting anatomical structures. This information is combined in a registration using diffeomorphic demons with a model that can assign individual regularisers to each of the anatomical objects.
- K (Lyon FFD) [55]: The lungs are firstly aligned with an affine registration. Secondly, the interface where sliding motion occurs is automatically segmented. Finally the detected interface is used to guide an intensity-based B-spline registration using mutual information as a similarity measure.
- L (MGH) [56], [57]: The images were masked using the provided segmentation results, and then translated to align the masks. Next, a multi-resolution B-spline transform was optimized with L-BFGS to minimize an SSD cost function.
- M (Nifty Reggers) [58], [59]: A block-matching technique was used to perform an initial affine alignment. It was followed by three non-rigid steps, firstly to coarsely align the lung features, followed by the borders and finally the details. The non-rigid registration was based on a cubic B-Splines model and was driven by the NMI.
- N (Oxford Flow Discontinuity Preserving) [60]: After a histogram-matching step, a computationally efficient optical-flow based variational registration is performed using SAD as a similarity measure. A modified Lp norm is used for a robust, discontinuity preserving regularisation of the deformations.
- O (Philips Research) [61]: A fully-automatic, volumetric, multi-resolution algorithm consisting of (1) an affine registration step and (2) a non-rigid, non-parametric registration step. The second step simultaneously minimizes the SSD and a regularising term based on the Navier-Lamé operator.
- P (picsl exp) [62], [63]: An affine alignment using lung masks is performed as a first step. This is followed by a deformable registration with local NCC as a simililarity metric and an exponential mapping model. The whole registration was implemented using the open source Advanced Normalization Tools (ANTS) software package.
- Q (picsl gsyn) [62], [63]: The registration pipeline begins with an affine alignment using lung masks, which precedes greedy symmetric normalization coupled with local NCC. The whole registration was implemented using the open source Advanced Normalization Tools (ANTS) software package.
- R (PVG) [64]: A pre-processing stage is used to subsample the original image volumes and dilate the lung masks. The pre-processed volumes are then registered with a B-spline deformation and by the optimization of a gradient-orientation based similarity metric.
- S (Robust TreeReg Leuven) [65]: This method is based on the spline MIRIT algorithm (T). Prior to the dense registration, vessel bifurcations are detected and matched. During dense registration, a penalty is added based upon the distance between these corresponding bifurcations.
- T (Spline MIRIT Leuven) [65]: A B-spline registration is adopted using MI as similarity measure and L-BFGS-B as an optimizer. A multi-resolution approach is used in relation to both the transformation field and the image size.

VII. Results

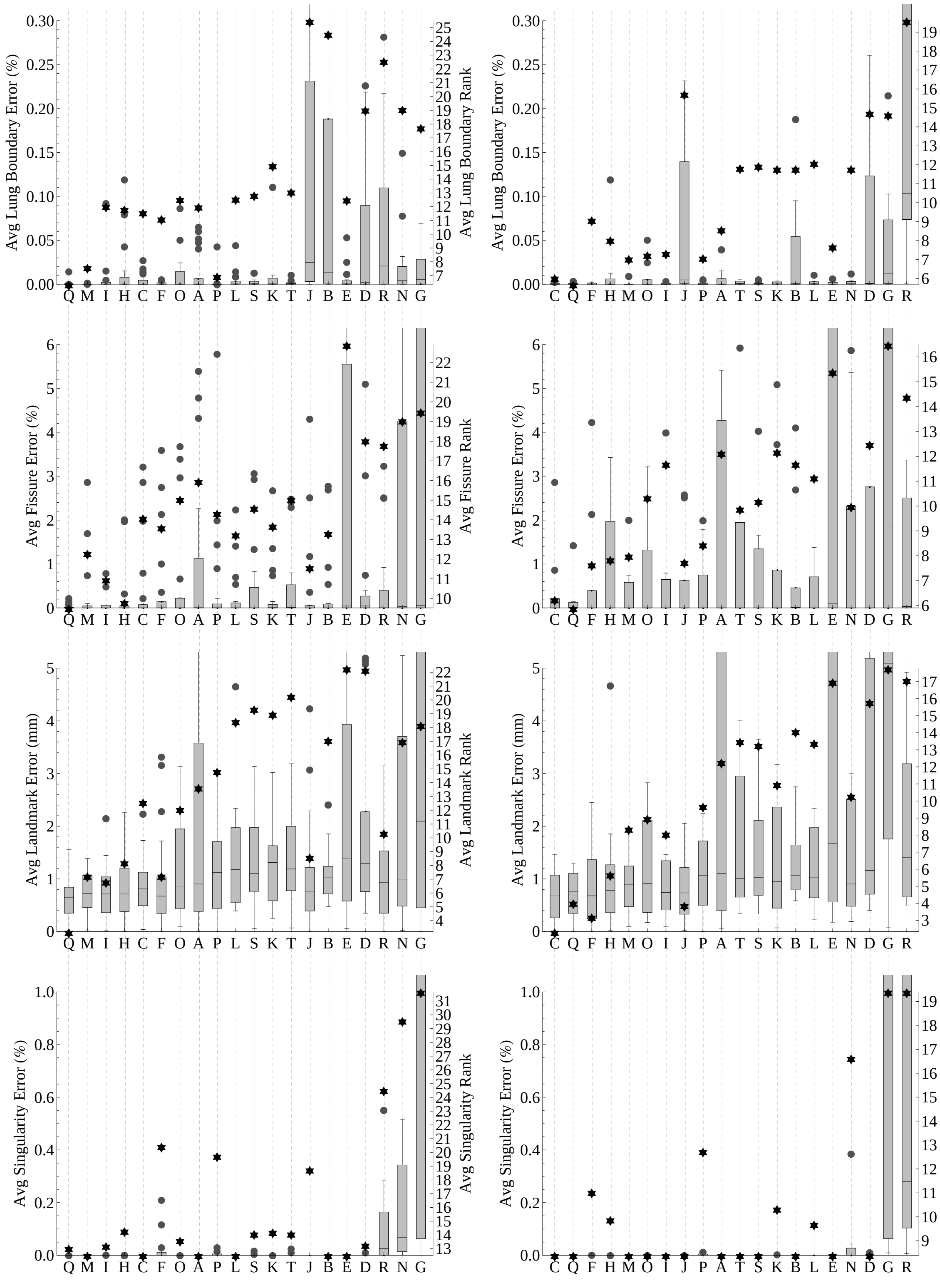
A. Phase 1 Results

Table IV gives the scores and ranks for each algorithm in each of the four categories, averaged over the 20 scan pairs that were registered. The algorithms are listed in order of their final placement in this phase. The overall average rank r for the algorithm, shown in the second last column, defines its final placement in this phase (last column). Figure 6(a) shows boxplots[2] illustrating the range of the error scores over the 20 scan pairs for each team and each category individually. The average ranking per category (as shown in table IV) is also plotted for reference.

B. Phase 2 (Workshop) Results

The scores and ranks for each algorithm in phase 2 are provided in table V. As with table IV the algorithms are listed

[2] Boxplots are used to represent the spread of values in a dataset. For all boxplots in this work the box spans from the 0.25 quantile to the 0.75 quantile with a horizontal line showing the median of the data. The whiskers are vertical lines that extend to indicate either the full dataset or the lowest/highest data point within 1.5 times the interquartile range. In the latter case remaining data points are plotted as outliers.



(a) (b)

Fig. 6. Boxplots showing the range of scores (errors) obtained in each category for each participant. (a) Phase 1 (20 scan pairs), (b) Phase 2 (10 scan pairs). Evaluation categories, from top to bottom, are: Lung Boundary Alignment, Fissure Alignment, Landmark Alignment, Singularity Scores. Participant labels are shown on the X-axis in order (left to right) of their final placement in that phase. The left Y-axis shows the score values. The * symbol at each boxplot represents the average ranking of the participant in that category, with scales shown on the right Y-axis. Note that the average ranking is based on the individual rankings per scan-pair and not on the average scores, therefore a linear relationship between average score and average ranking is not to be expected. Boxplot outliers are denoted by filled circles.

Label	Automatic?	Open Source?	Non-rigid Transformation Model	Similarity Measure	Lung Masks?	Placement Phase 1 (/34)	Placement Phase 2 (/20)
A	Fully	✗	Displacement field	SSD	✓	9	10
B	Fully	✗	Displacement field	CC	✓	17	14
C	Fully	✗	Dense displacement field	Hybrid MI/SSD	✓▷	6	1
D	Fully	✗	B-Spline	MPSSD	✓	20	17
E	Fully	✗∗†	B-Spline	SAD	✗	19	16
F	Fully	✓∗	B-spline	NCC	✓	7	3
G	Fully	✗	Optical Flow	SAD	✓	28	19
H	Fully	✗∗○	Diffeomorphic with static velocity fields	NSSD	✓	4	4
I	Fully	✗	B-Spline	SSTVD/ SSVMD	✓	3	7
J	Fully	✗	Diffeomorphic Diffusion	NSSD	✓	16	8
K	Fully	✓	B-Spline	MI	✓	14	13
L	Semi- (3)	✓∗	B-Spline	SSD	✓	12	15
M	Fully	✓∗	B-Spline	NCC/ NMI	✓	2	5
N	Fully	✗	Optical Flow	SAD	✗	26	17
O	Fully	✗	Non-parametric / Navier-Lame	SSD	✓	8	6
P	Fully	✓∗	Diffeomorphic (Exponential mapping)	NCC	✓	11	9
Q	Fully	✓∗	Diffeomorphic (Symmetric Normalization)	NCC	✓	1	2
R	Fully	✗	B-Spline	Adaptive LMI	✓	22	20
S	Fully	✗	B-Spline	MI	✗	13	12
T	Fully	✗	B-Spline	MI	✗	15	11

TABLE III

Summary of the algorithms entered in the EMPIRE10 challenge. Methods requiring no interaction and using the same parameters for all scan pairs are marked as **‘Fully’** automatic. Methods requiring more than one set of parameters for the list of 30 scan pairs are marked **‘Semi-’** automatic with the number of sets of parameters in brackets. **Open Source Symbols:** ∗ implies that the full set of parameters used is available either through the MICCAI publication or through the website where the software can be downloaded, enabling the reader to fully implement the registration described. † implies that the binaries are available for download although the code is not open source. ○ implies that the algorithm is intended to be made open source in the near future. The ‘Lung Masks’ column indicates whether binary lung segmentations were used during registration. ▷ implies that the lung masks were used in phase 2 only. Acronyms and abbreviations may be found in appendix A.

in order of their placement in this phase of the challenge. The scores and ranks shown are averaged over the 10 scan pairs processed in phase 2. The range of errors over the 10 scan pairs is shown for each team in each category in figure 6(b). As for the phase 1 data, the average rank for the category is plotted along with each box plot. Table VI provides additional important information in relation to the processing of the last 10 scan pairs. Since the majority (16 of the 20 algorithms) of registrations were computed during the 3 hour time slot at the MICCAI Grand Challenge Workshop it is important to make note of which algorithms were run at a later date (during the week following the workshop). Furthermore the table details whether the hardware used was on site at the workshop (laptops only) or remotely at the participant’s own institute, allowing for the possibility to use much greater computing power. Information regarding the hardware used by each algorithm, and the average time taken to process a scan pair are given. Finally, a number of participants made some alterations to their algorithms between phase 1 and phase 2. These were mainly to improve the speed of processing but occasionally also to improve registration performance. Any changes made are noted in the rightmost column of table VI.

VIII. Discussion

The high level of interest in the EMPIRE10 challenge emphasises the fact that non-rigid registration remains a very active research topic, and that researchers recognise the importance of evaluating their algorithms in a comparable and objective manner. Our aim in organising this challenge was not to find the ‘best’ algorithm for the task at hand, but rather to provide a useful platform for comparison. Although not all researchers involved are working specifically in the field of thoracic CT, applying their registration algorithm to the EMPIRE10 data set enables them to obtain a quantitative reproducible evaluation which can be updated at any time to reflect the latest improvements to their method.

The organisation of this challenge has been of great benefit not only to individual research groups who were able to assess their algorithm’s performance, but also to the registration community at large. In the remainder of this section we discuss the outcome of the challenge and what has been learned about registration evaluation, about the registration of thoracic CT in particular and about the state of the art in non-rigid registration.

	Lung Boundaries		**Fissures**		**Landmarks**		**Singularities**		**Overall**	
Label	**Avg Score**	**Avg Rank**	**Avg Score**	**Avg Rank**	**Avg Score**	**Avg Rank**	**Avg Score**	**Avg Rank**	**Avg Rank**	**Placed (/34)**
Q	0.00	6.37	0.03	9.52	0.66	3.20	0.00	13.05	8.03	1
M	0.00	7.62	0.27	12.30	0.75	7.25	0.00	12.52	9.92	2
I	0.00	12.05	0.08	10.97	0.79	6.85	0.00	13.22	10.77	3
H	0.01	11.85	0.22	9.82	0.89	8.22	0.00	14.32	11.05	4
C	0.00	11.60	0.47	14.10	0.91	12.60	0.00	12.52	12.70	6
F	0.00	11.15	0.50	13.62	0.99	7.27	0.02	20.47	13.13	7
O	0.01	12.55	0.56	15.07	1.24	12.10	0.00	13.62	13.33	8
A	0.01	12.02	1.38	15.97	2.47	13.67	0.00	12.52	13.55	9
P	0.00	6.97	0.53	14.35	1.29	14.82	0.00	19.77	13.98	11
L	0.00	12.60	0.26	13.27	1.40	18.45	0.00	12.52	14.21	12
S	0.00	12.87	0.47	14.62	1.43	19.35	0.00	14.10	15.23	13
K	0.00	15.02	0.30	13.70	1.35	19.00	0.00	14.22	15.48	14
T	0.00	13.10	0.49	15.07	1.48	20.29	0.00	14.12	15.65	15
J	0.10	25.50	0.42	11.60	1.12	8.62	0.00	18.77	16.12	16
B	0.29	24.55	0.36	13.32	1.13	17.10	0.00	12.52	16.87	17
E	0.00	12.52	2.48	22.92	3.03	22.30	0.00	12.52	17.56	19
D	0.04	19.07	0.99	18.04	2.19	22.20	0.00	13.30	18.15	20
R	0.09	22.60	0.54	17.82	1.10	10.37	0.09	24.55	18.83	22
N	0.01	19.10	2.48	19.07	2.26	17.00	0.24	29.60	21.19	26
G	0.01	17.77	2.87	19.50	4.56	18.17	3.03	31.67	21.78	28

TABLE IV

RESULTS FROM PHASE 1. THE ALGORITHMS ARE LISTED IN ORDER OF THEIR FINAL PLACEMENT IN THIS PHASE, FROM FIRST TO LAST. SCORES AND RANKS ARE AVERAGED OVER THE 20 SCAN PAIRS THAT WERE REGISTERED AND ARE ROUNDED TO 2 DECIMAL PLACES. MORE DETAILED INFORMATION INCLUDING THE PERFORMANCE OF EACH TEAM ON EACH SCAN PAIR CAN BE FOUND ON THE EMPIRE10 WEBSITE [19].

	Lung Boundaries		**Fissures**		**Landmarks**		**Singularities**		**Overall**	
Label	**Avg Score**	**Avg Rank**	**Avg Score**	**Avg Rank**	**Avg Score**	**Avg Rank**	**Avg Score**	**Avg Rank**	**Avg Rank**	**Placed (/20)**
C	0.00	6.05	0.11	6.25	0.59	2.34	0.00	8.39	5.76	1
Q	0.00	5.70	0.16	5.90	0.65	4.05	0.00	8.39	6.01	2
F	0.00	9.10	0.46	7.65	0.77	3.20	0.00	11.05	7.75	3
H	0.00	8.05	0.61	7.85	1.06	5.70	0.00	9.89	7.87	4
M	0.00	7.05	0.26	8.00	0.88	8.39	0.00	8.39	7.96	5
O	0.00	7.25	0.58	10.35	1.03	9.00	0.00	8.39	8.75	6
I	0.21	7.35	0.52	11.70	5.04	8.10	0.00	8.39	8.88	7
J	0.03	15.75	0.32	7.75	0.72	3.90	0.00	8.39	8.95	8
P	0.00	7.10	0.25	8.45	1.03	9.70	0.00	12.75	9.50	9
A	0.00	8.60	0.95	12.15	2.02	12.30	0.00	8.39	10.36	10
T	0.00	11.85	0.87	9.89	1.44	13.50	0.00	8.39	10.91	11
S	0.00	11.95	0.61	10.20	1.32	13.30	0.00	8.39	10.96	12
K	0.00	11.80	0.89	12.20	1.84	11.00	0.12	10.35	11.33	13
B	0.02	11.80	0.46	11.70	1.30	14.10	0.00	8.39	11.50	14
L	0.06	12.10	1.68	11.15	4.51	13.40	1.62	9.70	11.58	15
E	0.00	7.70	2.23	15.40	2.34	17.00	0.00	8.39	12.12	16
N	0.00	11.80	0.85	10.00	1.08	10.30	0.00	16.65	12.18	17
D	0.05	14.75	1.26	12.50	2.14	15.80	0.00	8.39	12.86	18
G	0.04	14.65	4.94	16.50	6.52	17.79	2.16	19.40	17.08	19
R	1.93	19.60	0.63	14.40	2.61	17.10	1.45	19.40	17.62	20

TABLE V

RESULTS FROM PHASE 2. THE ALGORITHMS ARE LISTED IN ORDER OF THEIR FINAL PLACEMENT IN THIS PHASE, FROM FIRST TO LAST. SCORES AND RANKS ARE AVERAGED OVER THE 10 SCAN PAIRS THAT WERE REGISTERED AND ARE ROUNDED TO 2 DECIMAL PLACES. MORE DETAILED INFORMATION INCLUDING THE PERFORMANCE OF EACH TEAM ON EACH SCAN PAIR CAN BE FOUND ON THE EMPIRE10 WEBSITE [19].

A. Categories of Evaluation

As described in section IV the EMPIRE10 challenge made use of 4 categories of evaluation, each weighted equally in determining the final placement of an algorithm. Using figure 6 as an illustration we consider the merit of each of these categories individually.

1) Singularities: Singularity assessment was included to ensure that registration results represented meaningful and physically plausible deformations. It can be seen from the average singularity scores given in tables IV and V as well as from the singularity score plots in figure 6 that very few of the algorithms had any significant problem with singularities

Label	At Workshop?	Hardware Local?	Hardware	# parallel cores per pair	GPU Used?	Avg time per pair (mins)	Method Changes post Phase 1
A	✓	✓	Laptop, Intel Core2 Duo @ 2.40GHz, 8GB RAM	2	✗	6.3	Multi-resolution scheme altered to improve speed
B	✓	✓	Laptop, Intel Core i5 (Quad core) @ 2.4GHz, 8GB RAM	4	✗	9.3	Block-matching within lungs only to improve speed
C	✓	✗	HP XW8400, Intel Xeon Quad-core @ 2.66GHz, 4GB RAM, NVIDIA GTX 280	4	✓	4.5	Added use of the lung masks to improve alignment
D	✗	-	Intel Xeon X5355, Quad core, 2 processors @ 2.66GHz, 16GB RAM	1	✗	112	-
E	✓	✓	Laptop, Intel Core2 Duo Mobile @ 2.16GHz, 4GB RAM	2	✗	1.5	-
F	✓	✓	Laptop, Intel Core i5 (Quad core) @ 2.5GHz, 4GB RAM	1	✗	18	-
G	✗	-	Intel Xeon E5540 (Quad core) @ 2.53GHz, Nvidia Tesla C1060, 4GB RAM	4	✓	4	Parameter changes
H	✓	✓	Intel Xeon (Quad core) @ 2.67GHz, 16GB RAM	4	✗	9	Multi-threading / CUDA used in pre-registration step
I	✓	✗	Dual processor Intel Xeon E5520 (Quad core) @ 2.27GHz, 48GB RAM	16	✗	77	B-Spline interpolation method altered
J	✓	✗	Intel QuadCore @ 2.66GHz, 8GB RAM	4	✗	15	-
K	✓	✗	Dual processor Intel Xeon (Quad-core) L5430 @ 2.66GHz, 16GB RAM	8	✗	45	B-spline grid resolution altered
L	✓	✓	Laptop, Intel Centrino Dual Core Duo T7500 @ 2.2GHz, 2GB RAM	2	✗	9	Parameter changes (2 pairs)
M	✓	✓	Laptop, Intel Core i7 Q720 @ 1.6GHz, 8GB RAM, NVidia Quadro FX 2800m	1	✓	5.5	-
N	✓	✓	Laptop Intel Duo Core @ 2.4 GHz, 4 GB RAM	2	✗	10	histogram matching step added
O	✓	✓	Laptop Intel Core2 Duo @ 2.66 GHz, 3.5GB RAM	1	✗	16	Stopping criterion added
P	✗	-	Intel Xeon E5450 2-core @ 3GHz, 16GB RAM	1	✗	230	-
Q	✓	✗	Intel Xeon E5450 2-core @ 3GHz, 16GB RAM	2	✗	69	Multi-threading added
R	✗	-	Dual processor Intel Q6700 (Quad-core) @ 2.66GHz, 8GB RAM	8	✗	40	Multi-resolution scheme altered to improve speed
S	✓	✗	Dual processor dual-core AMD Opteron 2220 @ 2.8 GHz, 32 GB RAM	1	✗	115	Stopping criterion added
T	✓	✗	Dual processor dual-core AMD Opteron 2220 @ 2.8 GHz, 32 GB RAM	1	✗	105	Stopping criterion added

TABLE VI

Information relating to the processing for phase 2 (workshop at MICCAI) for each algorithm. The second and third columns imply whether the processing was done during the morning of the workshop (or in the week following) and whether the participant used on-site (or remotely located) hardware if so. Columns 5, 6, and 7 list how many cores were used in parallel to perform each registration, whether or not GPU processing was used, and how long each registration took on average. The last column lists any algorithm or parameter changes made by the participant between phase 1 and phase 2.

in their deformations. In fact, many of them incorporated regularisation steps specifically to avoid any such issues. In each phase of the challenge, only 4 algorithms out of 20 had average singularity scores above 0 when rounded to 2 decimal places. The worst average singularity score obtained by an algorithm (in either phase) was 3.03, meaning that on average 3.03% of the voxel locations within the lung volume were penalised for having implausible deformations.

Although it is important to ensure that registration results are physically plausible, this evaluation category was, in general, not very useful in distinguishing between algorithms. In addition the ranking system used was somewhat unsuited to handling such negligible differences between algorithm scores. In figure 6(b), for example, it can be seen that while a perfect score of 0% in the singularity category generated a singularity ranking of 8.4, algorithm F received a ranking of 11.05 with an average singularity score of just 0.0002% (unrounded figures may be obtained on the algorithm's results page on the challenge website [19]). A very minor error could therefore have a disproportionate effect on the algorithm ranking.

2) Lung Boundary Alignment: In thoracic CT the lung boundary is among the most easily recognised high contrast regions and should therefore be relatively easy to align. Furthermore, in the EMPIRE10 challenge the participants were provided with lung masks which many teams used to assist their methods with aligning the lung boundaries correctly in the initial stages of registration. However, it may be envisaged that an algorithm spending a lot of effort on aligning internal structures such as vessels might inadvertently result in poorly aligned lung boundaries. In figure 6(b), for example, it can be seen that algorithm J performs very well in all categories with the exception of lung boundary alignment. More generally however, the majority of algorithms are well adapted to aligning the lung boundaries and figure 6 illustrates that in most cases the error is close to zero for all scan pairs. Of the 20 participating algorithms, 11 in phase 1 and 13 in phase 2 had zero error in this category when rounded to two decimal places (see tables IV and V).

3) Fissure Alignment: Fissure alignment was included as an evaluation category for three main reasons. Firstly, the fissures form important physical boundaries within the lungs, and therefore any algorithm which would be intended for use in a clinical application should be able to align them accurately. Secondly, the fissures are frequently difficult to register or even to detect, and therefore present an interesting challenge. Finally, the points used in the landmark category are rarely located on fissures so their alignment is not well evaluated in that category. Fissures are plate-like structures which are very narrow in one direction, and with the partial volume effect they are often comparatively low-contrast or, in poor quality data, partially obliterated by noise. Figure 7(a) shows an example of a fissure in an ultra-low-dose expiration scan that is moderately difficult to identify.

Tables IV and V and figure 6 show that there is much more variance in algorithm scores in the fissure alignment category than in either singularity or lung boundary categories. They are, therefore, useful for providing some distinction between algorithms where singularity scores and lung boundary scores may have been uniformly good. It can be seen in figure 6(a) that algorithm A, for example, performs extremely well in the singularities category and quite well also in the lung boundary

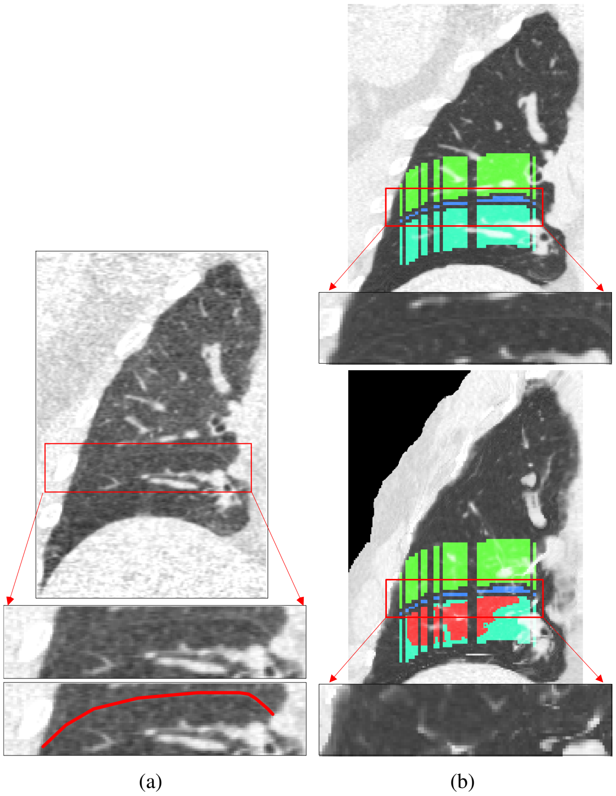


Fig. 7. (a) An example of a fissure as seen in an ultra low-dose expiration scan (fixed scan, pair 21). Although the fissure is visible in this scan, it has very low contrast and the image noise makes it more difficult to identify. The lower images show an enlarged view of the fissure and the same enlarged view with the fissure highlighted in red. (b) The deformed moving image showing the same slice as in figure a. Above: For the algorithm performing best on fissures for this scan pair (algorithm Q). Below: For the algorithm performing worst on fissures for this scan pair (algorithm G) (Note that the black region to the top left of this image implies only that deformation information was not supplied for that area since it is beyond the region of interest). The fissure reference standard is overlaid on both images. Red colouring is used to indicate locations where a penalty for fissure misalignment was incurred. The enlarged regions show that algorithm Q has deformed the fissure to the location specified by the reference standard whereas no fissure is visible in this region in the deformed image from algorithm G.

alignment category, with just a few close outliers. However in the fissure alignment category, although the median error remains close to zero, there are quite a few scan pairs for which performance was considerably worse, resulting in an extended boxplot and several distant outliers. In phase 2 most of the algorithms performed extremely well in terms of both singularities and lung boundary alignment, however differences are much more apparent in the fissure alignment category (see figure 6(b)). Examples of fissure alignment in scan pair 21 are shown in figure 7(b) for the algorithms which performed best and worst on fissures in this scan pair. This figure illustrates how penalties are incurred by the algorithm which failed to align the fissure correctly.

4) Landmark Alignment: Landmarks were included in the evaluation to give an insight into the ability of the algorithm to align small structures throughout the lung volume. Figure 6 illustrates that the landmark category was the best at distinguishing between registration results. The median values and box plot sizes are much more varied in this category than in any other, both for phase 1 and phase 2. Figure 8 shows a sample landmark from scan pair 21. The top row shows the landmark in the fixed scan, and three accepted observer opinions for the matching location in the moving scan. Subsequent rows show the matching location selected by each algorithm and its distance, d, to the closest observer choice. The distance values vary from 0 mm (perfect agreement with one of the observers) up to 67.1 mm. In fact scan pair 21 was one of the most difficult scan pairs to register due to a very large deformation between the inspiration and expiration scans, resulting in this diversity in algorithm results.

A total of 8 scan pairs were considered as special cases in terms of landmark evaluation since the point correspondence was not manually defined but known absolutely. These consisted of 4 scan pairs in which the images were related by artificial warping, and 4 scan pairs from ovine data where fiducial marker locations were known. Figure 9 compares average landmark error per participant for scan pairs where landmarks were manually annotated (x-axis) with landmark error on the warped and ovine cases (y-axis) respectively. In both scatter plots there is a reasonably good correlation (r=0.68 and r=0.75) between the average error scores, indicating that, generally speaking, algorithms which perform well on the artificial/fiducial data tend to perform well also on the manually annotated data. However the slopes, m of the lines fitted by least-squares reveal a disadvantage to the artificially warped data in particular. For the warped data the slope is 0.17, indicating that algorithms tend to have a very much lower average error value on the artificially warped data than on the manually annotated data. In the case of the ovine scans with fiducial markers, the slope of the line m=0.66 indicates that while the general trend is for a lower error in the ovine data compared to the manually annotated data, the distinction is much less obvious than in the case of the warped data. In fact it may be expected that the error in the manually annotated pairs would be slightly higher since these include the most difficult category of inspiration-expiration pairs (see section VIII-B). We conclude that the artificially warped data can be useful for comparison of performance between algorithms but it is not suitable for determining the actual accuracy that might be expected of an algorithm on real data. The fiducial markers, on the other hand, are useful both for comparing performance of algorithms and also in determining actual accuracy levels.

Based on the landmark category results from the best performing algorithms given in tables IV and V it may be suggested that there is very little room for improvement with average errors approximately equivalent to slice thickness being reported. However, although the average landmark distance over all scan pairs is excellent for these algorithms, the maximum landmark distance in each scan pair may not be so good, implying that there are small regions in the scan where alignment is incorrect. Table VII shows the maximum landmark distance d per scan pair in phase 2 for the best 3 algorithms in that phase (maximum distances for each algorithm and scan pair are available through the website results pages [19]). The overall average landmark score for each algorithm is shown in the last column for comparison. It can be seen that although an algorithm may have an excellent overall landmark score,

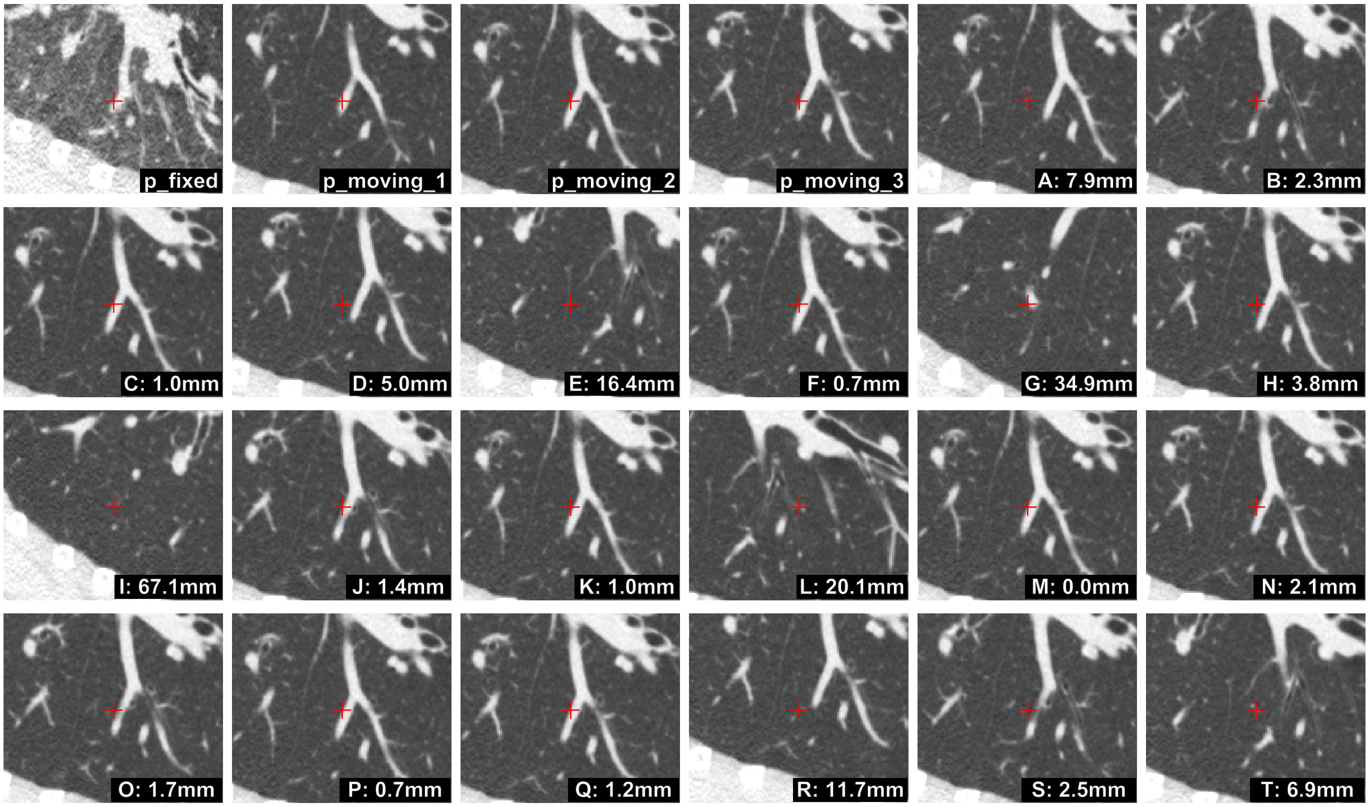
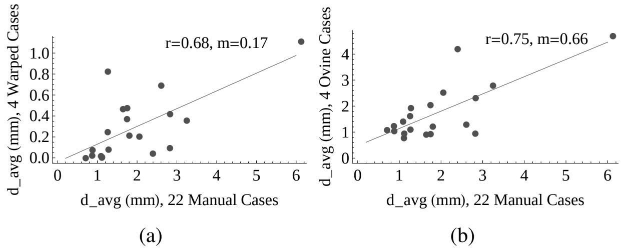


Fig. 8. A sample landmark point from scan pair 21. Top left: The location in the fixed scan and 3 (accepted) observer opinions about the matching location in the moving scan. The remaining images show the matching point chosen by each algorithm along with the distance d to the nearest observer chosen match. This particular landmark is shown because of the variety in the algorithm results due to scan pair 21 being among the most difficult data sets provided. All images shown are in the coronal direction.



(a)

(b)

Fig. 9. Scatterplots representing average landmark error, d_{avg}, in manually annotated cases compared to (a) cases where annotations were known due to artificial warping and (b) cases where annotations were given by fiducial marker locations in ovine data. Each plotted point corresponds to a participant. The x-axis value is the average of the landmark error scores, d_{avg} in the 22 manually annotated cases. The y-axis value is the average of the landmark error scores, d_{avg} in (a) 4 artificially warped cases and (b) 4 ovine data (fiducial markers) cases. The values r and m shown are, respectively, the correlation coefficient for the data and the slope of the line fitted by least squares.

there are still some landmarks in particular scan pairs which are not well aligned. Figure 10 shows an example of this for algorithm C and scan pair 28. The landmark where the algorithm performed worst ($d = 15.05mm$) is shown in this image, along with the point incorrectly chosen by algorithm C. It can be seen that although the majority of the scan is well aligned, there is a small region around this landmark where alignment is poor.

Pair	21	22	23	24	25	26	27	28	29	30	
Algorithm											[Avg]
C	6.04	6.53	2.86	4.67	1.23	1.65	1.28	15.05	8.87	0.00	[0.59]
Q	7.48	6.20	2.86	5.24	1.23	1.65	1.28	13.68	8.64	0.00	[0.65]
F	13.66	6.53	2.68	3.76	0.00	1.65	1.28	20.97	6.38	0.00	[0.77]

TABLE VII

THE MAXIMUM LANDMARK DISTANCE d PER SCAN-PAIR IN PHASE 2 FOR THE BEST 3 ALGORITHMS IN THAT PHASE. ALL DISTANCES ARE IN MM. THE LAST COLUMN SHOWS THE AVERAGE LANDMARK DISTANCE OVER ALL SCAN PAIRS (AS PER TABLE V) ILLUSTRATING THAT ALTHOUGH THE OVERALL AVERAGE MAY BE VERY LOW THERE ARE STILL SOME LANDMARKS IN SOME SCAN PAIRS WHICH ARE NOT WELL ALIGNED.

To further illustrate this point, figure 11 demonstrates the difference in performance when considering averages (i.e. overall landmark scores) compared to considering actual landmark distances with no averaging. In figure 11(a) the overall landmark scores (averages) are plotted, firstly with a maximum value of 5 on the y-axis, and secondly with the full range of y-values shown. (Note that the upper image in 11(a) is identical to the landmark image in figure 6(a)). In figure 11(b) all landmark error values are plotted individually without averaging over scan pairs. This shows a much larger number of outliers with a maximum outlier value of 53 mm. The method of averaging to achieve a final score per scan-pair, while convenient for comparison purposes, can be deceptive

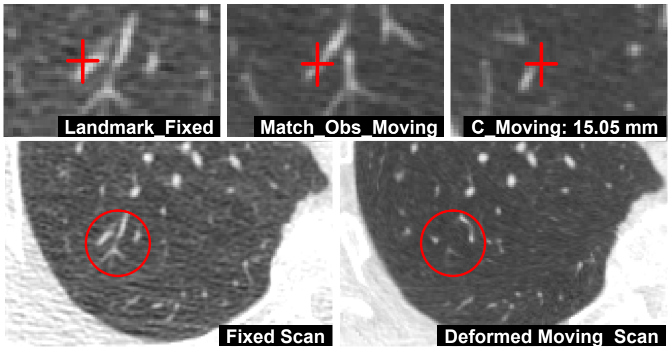
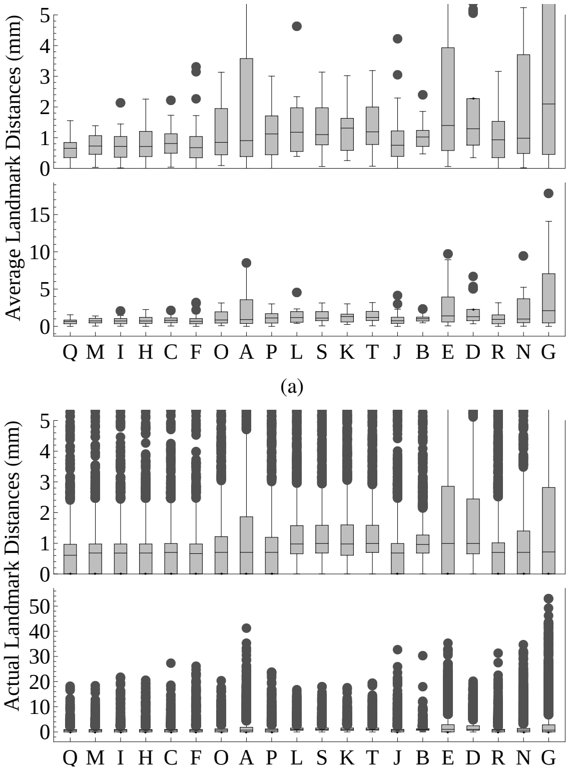


Fig. 10. A landmark in scan pair 28 which was incorrectly aligned by algorithm C. The top row shows the original landmark, the point chosen by an observer in the moving scan (closest observer point to algorithm C choice), and the point chosen by algorithm C in the moving scan, which was 15.05 mm away from the closest observer mark. The images in the second row show the fixed scan and the deformed moving scan according to algorithm C. The deformed scan aligns well with the fixed image in most locations, but close to the landmark (circled) the alignment is incorrect.



(a)

(b)

Fig. 11. (a) Boxplots showing the *average* landmark error per participant, where each value used is the average for a scan pair (20 values per plot). Shown with the y-axis range of [0-5] (above) and with the full range of y-axis values [0-18] (below). (b) Boxplots of the *actual* landmark errors per participant without averaging over each scan-pair first (1930 values per plot). Shown with the y-axis range of [0-5] (above) and the full range of y-axis values [0-53] (below). All data is from phase 1 of the challenge and participants are arranged in order of their placement in that phase.

when considering the individual performance of an algorithm. We can therefore conclude that even the best performing algorithms, although their average results are excellent, have some room for improvement in more difficult regions.

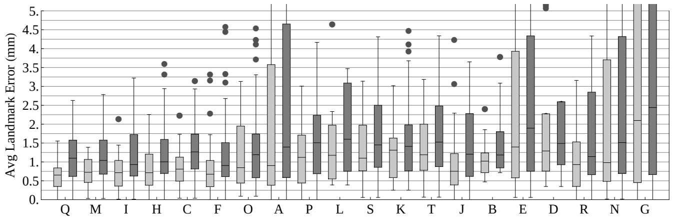


Fig. 12. Boxplots showing the landmark error per participant where the *closest* observer annotation was considered as the reference standard (light grey), and in contrast, the landmark error per participant if the *farthest* observer annotation had been considered as the reference standard (dark grey). All data is from phase 1 of the challenge and participants are arranged in order of their placement in that phase. Note that not the full range of data is shown.

Another issue to note is our choice to consider the observer point closest to the algorithm's location as the reference standard. In this way, if several observers have made different annotations, we opt to give the benefit of the doubt to the algorithm being evaluated. Since there can be a difference in the order of a few millimetres between observer marks it would be expected that all algorithms would disimprove in performance if we chose to define the reference standard in a different way. Figure 12 illustrates this point by showing the performance of algorithms according to the current reference standard compared with their performance if we use the *farthest* observer annotation as the reference standard. Since all observers are treated as equally correct this method of evaluation is just as valid as the method which is currently used. However it can be seen that performance is decreased, relatively severely in some cases, with median error values over the 20 scan pairs increasing by up to 0.5 mm and results at the upper whisker of the box plot increasing by several millimetres in some cases.

It must therefore be concluded that we confer some advantage to the performance of the algorithms in the landmark category by always using the closest observer point, and that actual performance may be somewhat poorer than reported. To fully resolve this issue however, would require knowledge of a single correct correspondence for every landmark, which in most cases is not feasible to determine.

B. *Categories of thoracic CT data*

As described in section II the thoracic CT pairs provided for the EMPIRE10 challenge came from a number of sources and had widely varying characteristics. The data was divided into 6 categories as follows: Inspiration-Expiration pairs (breath-hold), Inspiration-Inspiration pairs (breath-hold), pairs from 4D data sets, ovine data, artificially warped data and contrast-enhanced data.

Figure 13 shows the range of landmark error values obtained by the various algorithms for each scan pair, grouping the scan pairs into their data categories. The first category shown, inspiration-expiration was clearly the most difficult type of data. This category of data requires the largest deformations to resolve the registration since there is typically a considerable difference in lung volume between breath-hold inspiration and breath-hold expiration. A second contributary factor in the difficulty with registering these cases may be the ultra-low-dose protocol used in acquiring the expiration scans. This results in noisier data and since the inspiration scans are somewhat better quality, there may be small structures visible at inspiration which are not easy to detect in the expiration scan. Figure 13 shows that there is quite some variation within the inspiration-expiration category, with scan pair 08 being relatively easy to register and scan pair 21 being the most difficult. This variation is to be expected as the amount of deformation is very dependent on the subject's health and ability to breathe deeply as well as their regard for the instructions given during scanning. Furthermore scan quality varies depending on many factors such as the weight of the subject, movement during scanning etc.

The second most difficult data category appears to be the ovine data, although it is closely followed by the 4D and inspiration-inspiration pairs. The ovine data does not exhibit large deformations so is not expected to be particularly difficult to register. In some cases algorithms may have been tested and tuned on human data, and be less suited to this data type as a result. However, another likely cause for the larger landmark errors in these cases is the nature of the landmarks themselves. Since these landmarks are based on fiducial markers, and are not necessarily located on high contrast boundaries (see figure 5) there is less structure around them to guide the registration. Testing the algorithm behaviour at points that do not incorporate high contrast structures is likely to result in a drop in performance. Ideally, landmarks should be distributed throughout the parenchyma without regard to the structure or lack thereof, however in practice it is extremely difficult for a human observer to match points in low contrast regions with any degree of accuracy. Reference standards including low-contrast landmarks are therefore difficult to obtain.

The inspiration-inspiration and 4D data categories are approximately similar in terms of difficulty in registration. There is more variation among the inspiration-inspiration pairs, probably depending on whether the patient succeeded in the same level of breath-hold in both scans (taken several months apart), and whether precisely the same scanner settings were used. In the 4D category, one of the challenges for registration algorithms is to remain robust to artifacts, which are more commonly encountered in this type of data.

The artificially warped data provided relatively little challenge in most cases. Since the task was simply to resolve a thin-plate-spline warp, rather than the much more complicated motion associated with breathing, most algorithms performed well, in fact many algorithms obtained zero landmark error on these pairs. Similarly for the data pairs including a contrast-enhanced image, performance was very good. The contrast material did not present any difficulties, and since the scans were taken just 30 seconds apart there was virtually no motion to resolve.

C. Registration Algorithms Analysis

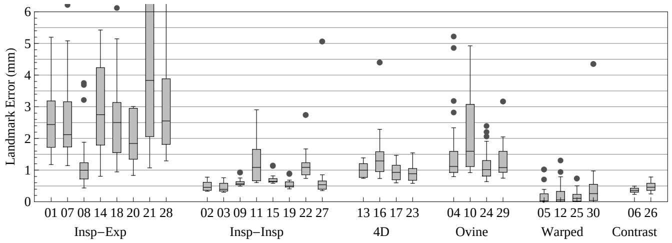


Fig. 13. Boxplots showing the landmark error per scan-pair for the range of participants. Outliers are denoted by filled circles (not the full range of data is shown). Scan pairs are grouped according to the type of data represented.

The competing algorithms in EMPIRE10 include a wide variety of registration types (transformation models, similarity measures etc.) as well as a selection of algorithms tailored towards thoracic CT applications and completely generic registration algorithms. All methods are fully automatic with the exception of one, (algorithm L) where parameters were manually altered for a few of the scan pairs. Many algorithms performed extremely well on many scan pairs and there is very little to choose between them. From table IV it can be seen that the top 6 algorithms all have average landmark distance scores of less than 1 mm, with a range from 0.66 mm to 0.99 mm. In phase 2, (see table V), the landmark scores for the top 6 algorithms ranged between 0.59 mm and 1.06 mm. When it is considered that locations are rounded to the nearest voxel before distance is determined and slice thickness is typically around 0.7 mm, these are extremely good results in spite of the errors remaining in some regions as described in section VIII-A4.

Considering the 5 algorithms which reached the top 3 in either phase 1 or phase 2 (algorithms C, F, I, M, Q), only 1 of these (algorithm I) was designed specifically for registration of thoracic CT data. In this case the similarity measures used included information about the tissue density between breathing phases and the 'vesselness' measure at each location. The remaining 4 algorithms are all generic registration methods which were applied to the EMPIRE10 data sets with appropriate parameter settings. It may be surmised, therefore, that at the present time and for this set of data, generic registration algorithms can perform just as well as, or better than, data specific methods. It may still be the case that combining aspects of both could improve performance even further, particularly on more difficult scan pairs.

The transformation models included among these 5 algorithms are B-Spline (three times), dense displacement field and a diffeomorphic transformation. Similarity measures are various forms of NCC, MI or SSD, with lung specific measures (SSTVD, SSVMD) used by algorithm I. These algorithm profiles are not notably different from others which performed less well in the challenge, therefore it may be concluded that the good performance of these methods is due to other more specific elements of the individual algorithms. It cannot be concluded that there is a single category of registration method which performs best on this type of data.

Since registration is evaluated only on the lung volume it seems logical that better results should be obtained by avoiding efforts to register structures outside the lungs. In fact, deformations in external regions may negatively impact the alignment of structures within the lung volume. Of the 20 participating algorithms, 16 of them, including the best performing methods among them, make use of lung segmentation masks in some way during registration (see table III). Algorithm C did not use the lung mask information during phase 1, but added a step to make use of it during phase 2 (see table VI). The improvement in its ranking from 6th place in phase 1 to 1st place in phase 2 is likely to be largely attributable to this alteration. Overall we conclude that the use of lung masks is to be recommended for optimal performance in registration of the lung volumes.

Regarding algorithm speed it is difficult to make generalisations since participants used their own hardware, which varied greatly (see table VI) and since not all algorithms are designed and programmed for optimal efficiency. With re-programming and better hardware the average time per scan pair might be very different for many methods. However, one point to note is that there is no evident trend of the best performing algorithms being the slowest. For example, algorithm C, which took first place in phase 2 of the challenge, took just 4.5 minutes per scan pair which was the third fastest of all algorithms. Therefore, there is every reason to be optimistic that excellent registration performance and efficiency which is acceptable in a clinical setting are not mutually exclusive traits.

D. Future Work

The EMPIRE10 challenge remains open to new or improved entries, thereby continually monitoring the current state of the art in registration of thoracic CT. In section VIII-A4 it was noted that although some of the best performing algorithms achieve excellent average landmark error scores, they still fail to align small regions of some scan pairs correctly. It is hoped that registration performance will continue to improve in the future, enabling correct alignment of these more difficult areas.

In spite of these outstanding issues, the standard of registration in the EMPIRE10 challenge is generally very high and there are some extremely accurate algorithms included among the participants. Depending on the clinical application in question, some of these may already be sufficiently good to aid medical personnel in their daily work. Additional challenges lie ahead in optimising the speed of algorithms to make them practical for use in a clinical setting, as well as embedding them into the workstations and daily routines of clinicians. Furthermore, while the current aim is to describe the patient motion in terms of the external coordinate system, an extremely interesting extension for the future would be to describe the breathing motion in terms of the patient's own coordinate system - a problem which is, as yet, relatively poorly defined [7]. However discussion of these objectives is beyond the scope of this work.

IX. Conclusion

The EMPIRE10 challenge has enabled detailed, independent and fair evaluation of non-rigid registration algorithms. Although the common data set was composed of intra-patient thoracic CT image pairs, generic algorithms which were not tailored for this data performed extremely well and many different approaches to registration were shown to be successful. The inspiration/expiration scan pairs proved to be the most difficult to register accurately because of the large deformations present. Among the most noteworthy conclusions reached in section VIII is that corresponding landmarks provide the most useful reference standard for distinguishing between registration algorithm results. Although such landmarks are typically tedious to obtain, a semi-automatic system [27], [28] for defining them was used in this work. It was also determined that the use of artificial warping as an evaluation method is beneficial in distinguishing between different algorithms, but not in providing a true evaluation of a particular algorithm's accuracy. Analysis based on fiducial markers in ovine images, however, was shown to give a good representation of algorithm accuracy as well as a means of comparing different methods.

The results of this challenge represent an important step forward, both for the non-rigid registration community and for those involved in bringing automatic processing into clinical practice. Researchers in registration may now evaluate their algorithms, and any methodological improvements applied to them, in a quantitative independent way. In addition, the state of the art in registration of thoracic CT has been established for the first time, enabling a logical analysis of what is required in the future to bring registration into the clinic.

Acknowledgements

The authors would like to thank the following people for their assistance in various aspects of organising the EMPIRE10 challenge: Marius Staring, Leiden University Medical Center, The Netherlands, Thessa Kockelkorn and Gerard van Hoorn, Image Sciences Institute, University Medical Center Utrecht, The Netherlands. Dataset *17* was provided courtesy of the Léon Bérard Cancer Center and CREATIS lab, Lyon, France. Datasets *13*, *16* and *23* are courtesy of the Department of Radiation Oncology, Stanford University School of Medicine and Philips Research.

Appendix
Acronyms and Abbreviations

- CC: Correlation Coefficient
- CUDA: Compute Unified Device Architecture
- FFD: Free Form Deformation
- L-BFGS (and the variant L-BFGS-B): Limited memory BFGS (Broyden Fletcher Goldfarb Shanno)
- LMI: Local Mutual Information
- MI: Mutual Information
- MPSSD: Mass Preserving Sum of Squared Differences
- MRF: Markov Random Field
- NCC: Normalised Cross Correlation
- NMI: Normalised Mutual Information
- NSSD: Normalised Sum of Squared Differences
- SAD: Sum of Absolute Differences
- SSD: Sum of Squared Differences
- SSTVD: Sum of Squared Tissue Volume Difference
- SSVMD: Sum of Squared Vessel Measurement Difference

References

[1] L. G. Brown, "A survey of image registration techniques," *ACM Computing Surveys*, vol. 24, pp. 325–376, 1992.

[2] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, pp. 1–36, 1998.

[3] H. Lester and S. R. Arridge, "A survey of hierarchical non-linear medical image registration," *Pattern Recognition*, vol. 32, pp. 129–149, 1999.

[4] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Physics in Medicine and Biology*, vol. 46, pp. R1–45, 2001.

[5] B. Zitová and J. Flusser., "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, pp. 977–1000, 2003.

[6] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[7] N. J. Tustison, T. S. Cook, G. Song, and J. C. Gee, "Pulmonary kinematics from image data: a review," *Academic Radiology*, vol. 18, no. 4, pp. 402–417, 2011.

[8] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.

[9] X. Gu, H. Pan, Y. Liang, R. Castillo, D. Yang, D. Choi, E. Castillo, A. Majumdar, T. Guerrero, and S. B. Jiang, "Implementation and evaluation of various demons deformable image registration algorithms on a GPU." *Physics in Medicine and Biology*, vol. 55, no. 1, pp. 207–219, Jan 2010.

[10] S. Kabus, T. Klinder, K. Murphy, B. van Ginneken, C. Lorenz, and J. P. W. Pluim, "Evaluation of 4D-CT lung registration." *MICCAI*, vol. 12, no. Pt 1, pp. 747–754, 2009.

[11] M. A. Yassa and C. E. L. Stark, "A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe." *Neuroimage*, vol. 44, no. 2, pp. 319–327, Jan 2009.

[12] D. Sarrut, B. Delhay, P. Villard, V. Boldea, M. Beuve, and P. Clarysse, "A comparison framework for breathing motion estimation methods from 4-D imaging." *IEEE Transactions on Medical Imaging*, vol. 26, no. 12, pp. 1636–1648, 2007.

[13] R. Castillo, E. Castillo, R. Guerra, V. E. Johnson, T. McPhail, A. K. Garg, and T. Guerrero, "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets." *Physics in Medicine and Biology*, vol. 54, no. 7, pp. 1849–1870, 2009.

[14] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer Jr., R. M. Kessler, R. J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P. A. van den Elsen, S. Napel, T. Sumanaweera, B. Harkness, P. F. Hemler, D. L. G. Hill, D. J. Hawkes, C. Studholme, J. B. A. Maintz, M. A. Viergever, G. Malandain, X. Pennec, M. E. Noz, G. Q. Maguire Jr., M. Pollack, C. A. Pelizzari, R. A. Robb, D. Hanson, and R. P. Woods, "Comparison and evaluation of retrospective intermodality image registration techniques," *Journal of Computer Assisted Tomography*, vol. 21, pp. 554–566, 1997.

[15] G. Christensen, X. Geng, J. Kuhl, J. Bruss, T. Grabowski, I. Pirwani, M. Vannier, J. Allen, and H. Damasio, "Introduction to the non-rigid image registration evaluation project (NIREP)." in *Third International Workshop on Biomedical Image Registration*, ser. Lecture Notes in Computer Science, vol. 4057, 2006, pp. 128–135.

[16] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. L. Goualher, D. L. Collins, A. Evans, G. Malandain, N. Ayache, G. E. Christensen, and H. J. Johnson, "Retrospective evaluation of intersubject brain registration." *IEEE Transactions on Medical Imaging*, vol. 22, no. 9, pp. 1120–1130, 2003.

[17] K. K. Brock and the Deformable Registration Accuracy Consortium, "Results of a multi-institution deformable registration accuracy study (MIDRAS)," *International Journal of Radiation Oncology Biology Physics*, vol. 76, no. 2, pp. 583–596, 2010.

[18] R. Kashani, M. Hub, J. M. Balter, M. L. Kessler, L. Dong, L. Zhang, L. Xing, Y. Xie, D. Hawkes, J. A. Schnabel, J. R. McClelland, S. Joshi, Q. Chen, and W. Lu, "Objective assessment of deformable image registration in radiotherapy: a multi-institution study." *Medical Physics*, vol. 35, no. 12, pp. 5944–5953, 2008.

[19] http://empire10.isi.uu.nl.

[20] http://www.grand-challenge.org/index.php/MICCAI_2010_Workshop.

[21] 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," *Medical Physics*, vol. 36, no. 7, pp. 2934–2947, 2009.

[22] 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*, vol. 54, no. 2, pp. 177–184, 2006.

[23] J. Vandemeulebroucke, D. Sarrut, and P. Clarysse, "The POPI model, a point-validated pixel-based breathing thorax model." in *XVth ICCR*, 2007.

[24] J. A. Schnabel, C. Tanner, A. D. Castellano-Smith, A. Degenhard, M. O. Leach, D. R. Hose, D. L. G. Hill, and D. J. Hawkes, "Validation of nonrigid image registration using finite-element methods: application to breast MR images." *IEEE Transactions on Medical Imaging*, vol. 22, no. 2, pp. 238–247, 2003.

[25] M. Urschler, S. Kluckner, and H. Bischof, "A framework for comparison and evaluation of nonlinear intra-subject image registration algorithms," in *IJ - 2007 MICCAI Open Science Workshop*, 2007.

[26] H. Wang, L. Dong, J. O'Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung, "Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy," *Physics in Medicine and Biology*, vol. 50, no. 12, pp. 2887–2905, 2005.

[27] K. Murphy, B. van Ginneken, J. P. W. Pluim, S. Klein, and M. Staring, "Semi-automatic reference standard construction for quantitative evaluation of lung CT registration." *MICCAI*, vol. 11, no. Pt 2, pp. 1006–1013, 2008.

[28] K. Murphy, B. van Ginneken, S. Klein, M. Staring, B. J. de Hoop, M. A.Viergever, and J. P. W. Pluim, "Semi-automatic construction of reference standards for evaluation of image registration," *Medical Image Analysis*, vol. 15, pp. 71–84, 2011.

[29] http://www.miccai.org.

[30] E. M. van Rikxoort, B. van Ginneken, M. Klik, and M. Prokop, "Supervised enhancement filters: application to fissure detection in chest CT scans," *IEEE Transactions on Medical Imaging*, vol. 27, no. 1, pp. 1–10, 2008.

[31] V. Boldea, G. C. Sharp, S. B. Jiang, and D. Sarrut, "4D-CT lung motion estimation with deformable registration: quantification of motion nonlinearity and hysteresis." *Medical Physics*, vol. 35, no. 3, pp. 1008–1018, 2008.

[32] I. D. Grachev, D. Berdichevsky, S. L. Rauch, S. Heckers, D. N. Kennedy, V. S. Caviness, and N. M. Alpert, "A method for assessing the accuracy of intersubject registration of the human brain using anatomic landmarks." *Neuroimage*, vol. 9, no. 2, pp. 250–268, 1999.

[33] E. Heath, D. L. Collins, P. J. Keall, L. Dong, and J. Seuntjens, "Quantification of accuracy of the automated nonlinear image matching and anatomical labeling (ANIMAL) nonlinear registration algorithm for 4D CT images of lung," *Medical Physics*, vol. 34, no. 11, pp. 4409–4421, 2007.

[34] A. Pevsner, B. Davis, S. Joshi, A. Hertanto, J. Mechalakos, E. Yorke, K. Rosenzweig, S. Nehmeh, Y. E. Erdi, J. L. Humm, S. Larson, C. C. Ling, and G. S. Mageras, "Evaluation of an automated deformable image matching method for quantifying lung motion in respiration-correlated CT images." *Medical Physics*, vol. 33, no. 2, pp. 369–376, 2006.

[35] T. Vik, S. Kabus, J. von Berg, K. Ens, S. Dries, T. Klinder, and C. Lorenz, "Validation and comparison of registration methods for freebreathing 4D lung CT," in *Proceedings of the SPIE*, vol. 6914, 2008, pp. 69 142P–69 142P–10.

[36] Z. Wu, E. Rietzel, V. Boldea, D. Sarrut, and G. C. Sharp, "Evaluation of deformable registration of patient lung 4DCT with subanatomical region segmentations." *Medical Physics*, vol. 35, no. 2, pp. 775–781, 2008.

[37] D. Rey, G. Subsol, H. Delingette, and N. Ayache, "Automatic detection and segmentation of evolving processes in 3D medical images: Application to multiple sclerosis." *Medical Image Analysis*, vol. 6, no. 2, pp. 163–179, 2002.

[38] V. Garcia, T. Vercauteren, G. Malandain, and N. Ayache, "Diffeomorphic demons and the EMPIRE10 challenge," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 91–98, 2010.

[39] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, Supp.1, pp. S61–S72, 2009.

[40] V. Garcia, O. Commowick, and G. Malandain, "A robust and efficient block-matching framework for non linear registration of thoracic CT images," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 137–146, 2010.

[41] X. Han, "Feature-constrained nonlinear registration of lung CT images," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 63–72, 2010.

[42] V. Gorbunova, P. Lo, H. Ashraf, A. Dirksen, M. Nielsen, and M. de Bruijne, "Weight preserving image registration for monitoring disease progression in lung CT," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, no. 5242, pp. 863–870.

[43] V. Gorbunova, J. Sporring, P. Lo, A. Dirksen, and M. de Bruijne, "Mass preserving image registration: Results of evaluation of methods for pulmonary image registration 2010 challenge," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 155–164, 2010.

[44] B. Glocker, N. Komodakis, N. Paragios, and N. Navab, "Non-rigid registration using discrete MRFs: Application to thoracic CT images," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 147–154, 2010.

[45] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios, "Dense image registration through MRFs and efficient linear programming," *Medical Image Analysis*, vol. 12, no. 6, 2008.

[46] M. Staring, S. Klein, J. H. C. Reiber, W. J. Niessen, and B. Stoel, "Pulmonary image registration with elastix using a standard intensity-based algorithm," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 73–79, 2010.

[47] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "elastix: a toolbox for intensity-based medical image registration." *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, Jan 2010.

[48] M. Urschler, M. Werlberger, E. Scheurer, and H. Bischof, "Robust optical flow based deformable registration of thoracic CT images," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 195–204, 2010.

[49] T. Pock, M. Urschler, C. Zach, R. Beichel, and H. Bischof, "A duality based algorithm for TV-L1-optical-flow image registration," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 4792, 2007, pp. 511–518.

[50] A. Schmidt-Richberg, J. Ehrhardt, R. Werner, and H. Handels, "Diffeomorphic diffusion registration of lung CT images," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 55–62, 2010.

[51] J. Ehrhardt, R. Werner, A. Schmidt-Richberg, and H. Handels, "Statistical modeling of 4D respiratory lung motion using diffeomorphic image registration," *IEEE Transactions on Medical Imaging*, vol. 30, no. 2, pp. 251–265, 2011.

[52] K. Cao, K. Du, K. Ding, J. M. Reinhardt, and G. E. Christensen, "Regularized nonrigid registration of lung CT images by preserving tissue volume and vesselness measure," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 43–54, 2010.

[53] K. Cao, K. Ding, G. E. Christensen, M. L. Raghavan, R. E. Amelon, and J. M. Reinhardt, "Unifying vascular information in intensity-based nonrigid lung CT registration," in *Biomedical Image Registration: 4th International Workshop, WBIR2010*, ser. Lecture Notes in Computer Science, vol. 6204, 2010, pp. 1–12.

[54] S. E. A. Muenzing, B. van Ginneken, and J. P. W. Pluim, "Knowledge driven regularization of the deformation field for PDE based non-rigid registration algorithms," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 127–136, 2010.

[55] J. Vandemeulebroucke, S. Rit, J. Schaerer, and D. Sarrut, "Deformable image registration with automated motion-mask extraction," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 119–125, 2010.

[56] G. C. Sharp, M. Peroni, R. Li, J. Shackleford, and N. Kandasamy, "Evaluation of plastimatch B-Spline registration on the EMPIRE10 data set," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 99–108, 2010.

[57] J. A. Shackleford, N. Kandasamy, and G. C. Sharp, "On developing B-spline registration algorithms for multi-core processors," *Physics in Medicine and Biology*, vol. 55, no. 21, pp. 6329–6351, 2010.

[58] M. Modat, J. R. McClelland, and S. Ourselin, "Lung registration using the NiftyReg package," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 33–42, 2010.

[59] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin, "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine*, vol. 98, no. 3, pp. 278–284, 2010.

[60] M. P. Heinrich, M. Jenkinson, M. Brady, and J. Schnabel, "Discontinuity preserving regularisation for variational optical-flow registration using the modified Lp norm," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 185–194, 2010.

[61] S. Kabus and C. Lorenz, "Fast elastic image registration," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 81–89, 2010.

[62] G. Song, N. Tustison, B. Avants, and J. C. Gee, "Lung CT image registration using diffeomorphic transformation models," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 23–32, 2010.

[63] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[64] D. De Nigris, D. L. Collins, and T. Arbel, "Deformable registration of chest CT scans with adaptive local mutual information," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 175–184, 2010.

[65] D. Loeckx, D. Smeets, J. Keustermans, J. Hermans, F. Maes, D. Vandermeulen, and P. Suetens, "3D lung registration using splineMIRIT and robust tree registration (RTR)," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 109–117, 2010.