Evaluation of Methods for Pulmonary Image Registration: The EMPIRE10 Study

Keelin Murphy[1], Bram van Ginneken[2], Joseph M. Reinhardt[3], Sven Kabus[4], Kai Ding[3], Xiang Deng[5], and Josien P.W. Pluim[1]

[1] University Medical Center, Utrecht, The Netherlands
[2] Radboud University Nijmegen Medical Centre, The Netherlands
[3] Department of Biomedical Engineering, The University of Iowa
[4] Philips Research, Hamburg, Germany
[5] Corporate Technology, Siemens Ltd., China

Abstract. EMPIRE10 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. Participants download 20 datasets from the internet, register them, and return the results for independent evaluation. Evaluation is carried out in four separate categories and participants are ranked according to their performance. All results are published on the EMPIRE10 website [1]. The Grand Challenge workshop at MICCAI 2010 [2] brings participants together to register a further 10 scan pairs live on site and for discussion and collaboration opportunities.

1 Introduction

For many years researchers have worked on registration algorithms for medical imaging applications [3, 7, 10, 11, 14, 24]. One such application is the alignment of intra-patient thoracic CT images, in particular of the lung and its associated structures. The lungs are highly deformable organs making accurate registration of them a challenging task requiring a non-rigid registration approach. However there are many scenarios in which intra-patient pulmonary registration is clinically useful. Registration of follow-up (temporally distinct) breathhold inspiration scans makes 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 may be useful for example in understanding the effects of a particular 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 [5, 8, 9, 23]. 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. There have been some initiatives in the past which provided common datasets and evaluation methods for the evaluation of registrations of brain images [4, 6, 21], while allowing the users to configure and run their own registration algorithms on the data. Brain registration is, however, quite a different topic to pulmonary registration, primarily because of the much smaller deformations which are encountered in brain data, many of which can be resolved with a rigid or affine registration.

The EMPIRE10 (Evaluation of Methods for Pulmonary Image REgistration 2010) challenge [1] described in this article provides a public platform for fair and meaningful comparison of registration algorithms applied to thoracic CT data. This challenge, organised in conjunction with the Grand Challenge workshop at MICCAI 2010 [2], invites participants to download a set of 20 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 and 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 presence of singularities (folding or tearing). Evaluation results are available on the EMPIRE10 website [1]. At the MICCAI workshop participants are requested to register a further 10 scan pairs live on site within 3 hours. The workshop session is also designed to foster discussion and collaboration between researchers involved in this field.

2 Materials

The materials for this challenge were gathered from a variety of 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 inspiration, full expiration, phase from 4D breathing data). Subjects may exhibit lung disease or appear healthy. 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 20 scan pairs registered by the participants in their own facilities. The remaining 10 scan pairs to be registered live at the MICCAI workshop are from the same range of sources and have a similar variety of properties. Each scan pair is taken from a single subject, i.e. only intra-patient registration is considered in this challenge. 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 remainder of this section describes the 20 scan pairs categorised by type. The IDs of the scan-pairs, which are provided when downloading, are supplied in italics.

2.1 Breathhold Inspiration Scan Pairs

Six of the twenty scan pairs (*02,03,09,11,15,19*) consisted of two breathhold inspiration scans. These scans were made as part of the Nelson Study [22]. In these 6 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 1mm with slice-spacing of 0.70mm. Pixel spacing in the X and Y directions varied from 0.68mm to 0.78mm with an average of 0.73mm.

2.2 Breathhold Inspiration and Expiration Scan Pairs

A further 6 scan pairs (*01,07,08,14,18,20*), 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. 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 1mm and slice spacing of 0.70mm. Pixel spacing in the X and Y directions varied from 0.63mm to 0.77mm with an average value of 0.70mm.

2.3 4D Data Scan Pairs

Three of the scan-pairs (*13,16,17*) (excluding the ovine data - see section 2.4) 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. The first two of the scan pairs were from a GE Discovery ST multislice PET/CT scanner while the third [19] was from a Philips Brilliance CT 16 Slice scanner. The 2 4D scans from the GE scanner had a total dose of 100mAs each, while the third scan had a total dose of 400mAs. Since each scan pair came from a 4D dataset the spacing was identical for the two scans in the pair. Slice-spacing was 1.25mm, 2.50mm and 2.00mm for the three scan pairs respectively, with pixel spacing in the X and Y directions at 0.98mm in all cases.

2.4 Ovine Data Scan Pairs

Two scan pairs (*04,10*) were ovine (sheep) data from a single 4D dataset. Sixty-seven metallic markers, 1.40mm 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. The first scan pair consists of the scan taken at airway pressure 8 cm H_2O paired with the scan at airway pressure 16 cm H_2O, while the second scan pair is made up of the scan taken at airway pressure 8 cm H_2O paired with the scan at airway pressure 24 cm H_2O. Slice spacing was 0.60mm with in-plane pixel spacing of 0.47mm in both directions.

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 3.3) .

2.5 Contrast - Non-Contrast Scan Pairs

One pair of scans (*06*) was used in which contrast material was present in one of the scans but not in the other. These scans were both made on a Siemens SOMATOM Sensation CT 64-slice scanner. The contrast scan (arterial phase) was acquired approximately 30 seconds after the non-contrast scan. Slice spacing was 1.50mm with pixel spacing in the X and Y directions of 0.60mm.

2.6 Artificially Warped Scan Pairs

Since registration algorithms are difficult to evaluate in a quantitative way, a frequently employed method (e.g. [15,16,20]) 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 2 scan-pairs (*05,12*) 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 breath-hold inspiration scans from the Nelson Study [22] were 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 [12, 13]. A thin-plate-spline model was created using the 100 pairs of matching points. Using this thin-plate-spline model 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 also naturally somewhat 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. Regions around the

edge of the warped image where no data values could be assigned were cropped away. The scan pair distributed to the challenge participants consisted of the original baseline scan and the artificially warped version of this scan.

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

3 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 [1].

3.1 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. Analysis is therefore 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. [17]. Lung segmentations were checked and altered manually where necessary. 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 left lung was masked out by a sphere centred at the centre-of-mass of both lungs combined. The radius of this sphere was defined by the Euclidean distance from the centre-of-mass of both lungs combined to the centre-of-mass of the left lung. The mediastinal region of the right lung was masked out in a similar manner and all voxels within either sphere are excluded from further processing.

Next, points within 20mm of the lung boundary were marked, excluding those within 2mm of the boundary to allow for minor inaccuracies in the lung segmentation. 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 1(a) as an example.

Each participant submits deformation field data for each registration carried out. Using this data, it is possible to calculate for each point p_{fixed} marked with

v_{in} or v_{out} in the fixed image, which point p_{reg} in the moving image has been aligned with this location. If p_{fixed} is marked with v_{in} and p_{reg} is marked with v_{out} then a unit penalty is incurred. Similarly the reverse situation where p_{fixed} is marked with v_{out} and p_{reg} is marked with v_{in} also incurs a unit penalty.

Error in lung boundary alignment is calculated as the percentage of checked points for which penalties were incurred. This value is given as the the overall score in the lung boundary alignment category. For information, the errors in the left lung, right lung, upper lung and lower lung are also calculated and displayed on the participant's results page on the challenge website [1].

3.2 Alignment of Major Fissures

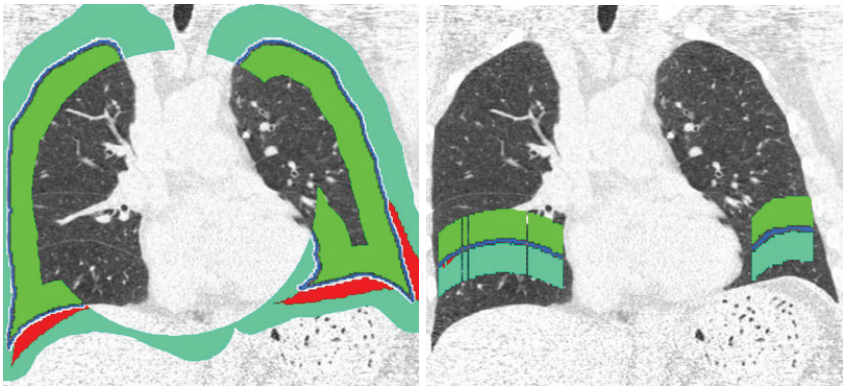
Fissures are plate-like structures which divide the lungs into regions called lobes. Since fissures represent important physical boundaries within the lungs their registration is included as a category in the EMPIRE10 challenge. To simplify the evaluation, particularly for poor quality data where minor 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.

The fissures in all images were segmented using an automatic algorithm from van Rikxoort et al. [18]. 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. A distance transform image was generated from the resulting fissure segmentation image.

Next, points within 20mm of the fissure segmentation were marked, excluding those within 2mm 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. 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 1(b) as an example.

Using the deformation data submitted by the participant, it is 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 has been aligned with this location. If p_{fixed} is marked with v_{above} and p_{reg} is marked with v_{below} then a unit penalty is incurred. Similarly the reverse situation where p_{fixed} is marked with v_{below} and p_{reg} is marked with v_{above} also incurs a unit penalty.

Error in fissure alignment is calculated as the percentage of checked points for which penalties were incurred. This value is given as the overall score in the fissure alignment category. For information, the errors in the left lung and right lung are also calculated and displayed on the participant's results page on the challenge website [1].



(a) (b)

Fig. 1. a: Lung boundary reference standard. The boundary itself is marked in blue and surrounded by a 2mm gap. Regions outside the lung are marked in cyan, and inside the lung are marked in green. Red regions indicate locations where a sample registration algorithm has made errors in aligning the lung boundary. **b:** Fissure reference standard. Colour coding is analogous to that in the left hand image with regions above and below the fissure marked in green and cyan respectively.

3.3 Correspondence of Annotated Landmark Pairs

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 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. [12, 13]. The software used is publicly available at http://isimatch.isi.uu.nl An example of the point distribution is shown in figure 2(a). The points were marked by at least 3 observers independently, and any location where the observer opinions differed by 3mm 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 rejected points were not included in the reference standard, all other points were retained. 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 is used to calculate for each of the defined points p_{fixed} in the fixed image which point p_{reg} in the moving image has been aligned with this location. The point p_{reg} is then compared (using Euclidean distance) with the reference standard point p_{moving}. Where several acceptable options for p_{moving} are defined, the p_{moving} that is closest to p_{reg} is used as the reference. Note that p_{reg} was rounded to the nearest voxel location before distance calculation. Since all observer marks are made without

sub-voxel accuracy this ensures that an algorithm which agrees precisely with a particular observer can obtain an error of zero.

The distance d from p_{moving} to p_{reg} is calculated in mm for each of the annotated point pairs. The overall error in the landmarks category is 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 lungs and in the lower lungs are also calculated and displayed on the participant's results page on the website [1]. The average distance in each of the three orthogonal directions (Anterior-Posterior, Superior-Inferior and Left-Right) are also calculated and displayed.

There are a few 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 2.4 and not manually annotated as for the other data. Furthermore, for the artificially warped data (see section 2.6) 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.

3.4 Singularities in the Deformation Field

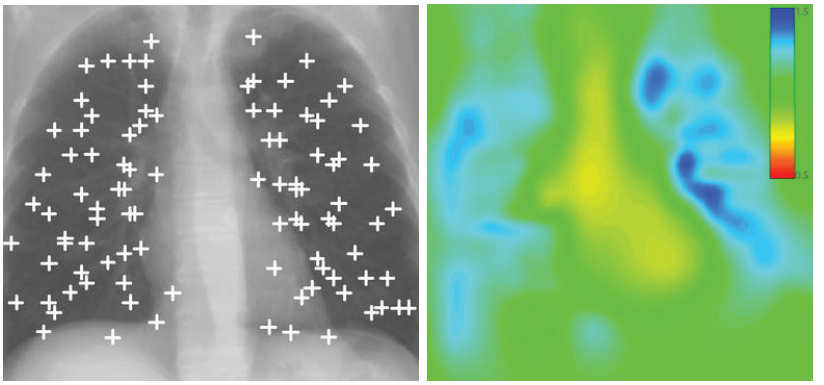
The final category of evaluation is designed to analyse how physically plausible the registration deformation is. Some registration algorithms may appear to align visible structures very well, but in doing so may require physically impossible deformations. In particular we expect that a deformation should 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 (folding or tearing).

Each participant submits deformation field data for each registration carried out. The determinant of the Jacobian of the deformation field, j, is calculated at every point. This specifies for each point whether local expansion or contraction has taken place. Where $j < 1$ local contraction is implied, $j = 1$ implies no change and $j > 1$ implies local expansion. Figure 2(b) shows an example of a colour coded Jacobian image. All points within the lung volume are checked and any location where $j \leq 0$ is a singularity in the deformation field. For each such point a unit penalty is incurred. Points outside the lung volume are disregarded.

The overall error in the singularities category is 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 are also calculated and reported on the participant's results page on the website [1].

4 Scoring and Ranking

Error scores in the four individual categories are calculated as described in section 3. A score is awarded to each participant for each scan-pair in each category.



(a) (b)

Fig. 2. a: An example of the landmark points identified in a fixed scan. Landmarks have been projected onto a single slice and markers are increased in size for visualization. **b:** A colour coded Jacobian image with the scale going from 0.5 (red) to 1.5 (blue). In this example there are no pixels at or below 0, so no singularities were detected.

Since these scores are based on independent measures 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 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 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: For simplicity assume that distinct rankings have been initially assigned, with participants with equal scores obtaining adjacent rankings (randomly assigned among them). 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 can be averaged over all scan-pairs to give each participant an average ranking r_c per evaluation category. 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 for rankings.

5 Results and Discussion

Results from a total of 34 algorithms from 23 research groups were submitted for the first 20 scan-pairs. Two algorithms were classed as semi-automatic while the remainder were fully automatic. The algorithms were evaluated and ranked as described in previous sections and the results published on the EMPIRE10 website [1]. Very few algorithms had noteworthy problems with singularities in the deformation field, while most also scored well in the alignment of lung boundaries. Fissure alignment appeared to be somewhat more difficult, with even the best algorithms showing minor errors in many of the scan pairs. Evaluation on the corresponding landmarks gave some error in almost every case. However, some algorithms achieved a perfect score for landmark matching in cases *05* and *12* which were the artificially deformed datasets. This implies either that the registration task was easier for those cases since the deformation was artificial, or that the reference standard being more precise improved the algorithm scores, or most likely, a combination of these two factors.

The degree of interest in this challenge emphasises the fact that non-rigid registration remains a very active research topic and that researchers continue to have difficulties in evaluating the performance of their algorithms. By providing this platform for evaluation we hope to have advanced understanding and enabled development in this field of research.

6 Acknowledgements

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

References

1. http://empire10.isi.uu.nl.
2. http://www.grand-challenge.org/index.php/MICCAI_2010_Workshop.
3. L. G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24:325–376, 1992.
4. G.E. Christensen, X. Geng, J.G. Kuhl, J. Bruss, T.J. Grabowski, I.A. Pirwani, M.W. Vannier, J.S. Allen, and H. Damasio. Introduction to the non-rigid image registration evaluation project (NIREP). In *Third International Workshop on Biomedical Image Registration. LNCS*, volume 4057, pages 128–135, 2006.
5. 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. *Phys Med Biol*, 55(1):207–219, Jan 2010.

6. P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D. L. Collins, A. Evans, G. Malandain, N. Ayache, G. E. Christensen, and H. J. Johnson. Retrospective evaluation of intersubject brain registration. *IEEE Trans Med Imaging*, 22(9):1120–1130, Sep 2003.
7. D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes. Medical image registration. *Phys Med Biol*, 46:R1–45, 2001.
8. S. Kabus, T. Klinder, K. Murphy, B. van Ginneken, C. Lorenz, and J.P.W. Pluim. Evaluation of 4D-CT lung registration. *MICCAI*, 12(Pt 1):747–754, 2009.
9. 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*, 46(3):786–802, Jul 2009.
10. H. Lester and S. R. Arridge. A survey of hierarchical non-linear medical image registration. *Pattern Recognition*, 32:129–149, 1999.
11. J. B. A. Maintz and M. A. Viergever. A survey of medical image registration. *Med Image Anal*, 2:1–36, 1998.
12. 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*, In Press, 2010.
13. 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*, 11(Pt 2):1006–1013, 2008.
14. J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Trans Med Imaging*, 22(8):986–1004, 2003.
15. J. A. Schnabel, C. Tanner, A. D. Castellano-Smith, A. Degenhard, M. O. Leach, D. Rodney Hose, D. L. G. Hill, and D. J. Hawkes. Validation of nonrigid image registration using finite-element methods: application to breast MR images. *IEEE Trans Med Imaging*, 22(2):238–247, 2003.
16. 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.
17. E. M. van Rikxoort, B. 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*, 36(7):2934–2947, 2009.
18. E. M. van Rikxoort, B. van Ginneken, M. Klik, and M. Prokop. Supervised enhancement filters: application to fissure detection in chest CT scans. *IEEE Trans Med Imaging*, 27(1):1–10, 2008.
19. J. Vandemeulebroucke, D. Sarrut, and P. Clarysse. The POPI model, a point-validated pixel-based breathing thorax model. In *XVth ICCR*, 2007.
20. H. Wang, L. Dong, J. O'Daniel, R. Mohan, A. S. Garden, K. Kian 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. *Phys Med Biol*, 50(12):2887–2905, Jun 2005.
21. 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. *J. Comp. Assist. Tomogr.*, 21:554–566, 1997.
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*, 54(2):177–184, 2006.
23. 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*, 44(2):319–327, Jan 2009.
24. B. Zitová and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, 2003.