Medical Image Analysis 15 (2011) 71–84

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Semi-automatic construction of reference standards for evaluation of image registration

K. Murphy *, B. van Ginneken, S. Klein, M. Staring, B.J. de Hoop, M.A. Viergever, J.P.W. Pluim

Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 26 May 2009
Received in revised form 20 July 2010
Accepted 20 July 2010
Available online 3 August 2010

Keywords:
Evaluation
Image
Registration
Validation
Ground-truth

ABSTRACT

Quantitative evaluation of image registration algorithms is a difficult and under-addressed issue due to the lack of a reference standard in most registration problems. In this work a method is presented whereby detailed reference standard data may be constructed in an efficient semi-automatic fashion. A well-distributed set of n landmarks is detected fully automatically in one scan of a pair to be registered. Using a custom-designed interface, observers define corresponding anatomic locations in the second scan for a specified subset of s of these landmarks. The remaining $n - s$ landmarks are matched fully automatically by a thin-plate-spline based system using the s manual landmark correspondences to model the relationship between the scans. The method is applied to 47 pairs of temporal thoracic CT scans, three pairs of brain MR scans and five thoracic CT datasets with synthetic deformations. Interobserver differences are used to demonstrate the accuracy of the matched points. The utility of the reference standard data as a tool in evaluating registration is shown by the comparison of six sets of registration results on the 47 pairs of thoracic CT data.

1. Introduction

Medical image registration is a well-established field of research upon which a substantial body of literature has been built over the past three decades (Brown, 1992; Maintz and Viergever, 1998; Lester and Arridge, 1999; Hill et al., 2001; Zitová and Flusser, 2003; Pluim et al., 2003). Although there have been significant advancements in registration techniques themselves, relatively little consideration has been given to methods of registration evaluation. Visual assessment may be sufficient to distinguish a very poor registration from an excellent one, particularly for 2D images. However, as more sophisticated algorithms are developed the distinction between the resulting registrations becomes more subtle and difficult to assess by eye. In the case of 3D and 4D image registrations visual assessment becomes unfeasible and alternative quantitative measures of registration accuracy are required.

Registration techniques may be divided into rigid and non-rigid classes. Rigid registrations involve rotation and translation at the most basic level, or in the case of affine transformations, may also include scaling and shearing. Non-rigid registrations handle elastic or fluid deformations, which must model far more complex motion, typically with many local changes in direction and magnitude of deformation. Since rigid registrations have far fewer degrees of freedom they represent a class of problem which is simpler to solve, and for which results are easier to evaluate. For example, by the manual specification of just a few corresponding locations, the desired transformation may be defined. For non-rigid registration however, both the solution and the evaluation of proposed solutions are highly complex problems.

The major obstacle in quantitative evaluation of non-rigid registration algorithms is the lack of any reference standard. The desired image transformation is rarely, if ever, known and manual specification of the full transform is simply not possible. The literature on registration evaluation methods may, in general, be divided into methods which attempt to evaluate in the absence of a reference standard (Glatard et al., 2006; Schnabel et al., 2003; Urschler et al., 2007; Wang et al., 2005), and methods which rely on a defined reference standard, which is unobtainable for most researchers (Škerl et al., 2008; Crum et al., 2004; Grachev et al., 1999; Heath et al., 2007; Blaffert and Wiemker, 2004; Betke et al., 2003; Vik et al., 2008; Castillo et al., 2009; Wu et al., 2008; Boldea et al., 2008; Pevsner et al., 2006).

A number of authors propose investigating registration performance by synthetically warping data such that the original image and the transformed image are known in advance as well as the ideal transform between them (Schnabel et al., 2003; Urschler et al., 2007; Wang et al., 2005). No reference standard needs to be defined for this type of evaluation, since it is already implied by the synthetic transformation used. Although attempts are made

* Corresponding author.
E-mail address: keelin@isi.uu.nl (K. Murphy).

1361-8415/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2010.07.005

to use physically plausible transformations, this approach provides only a generic evaluation and an algorithm's performance on real clinical data cannot be measured in this way. Glatard et al. (2006) have described a method of estimating the ground-truth with a 'bronze standard' obtained through the application of many registration algorithms to a large database of images. However, the method has not been extended to non-rigid registration problems and is very dependent on having a large number of good registration algorithms as well as an extensive database, both of which are difficult for many researchers to obtain.

Another common method of non-rigid registration evaluation is to measure the overlap of structures of interest in the target and registered images (Crum et al., 2004; Hellier et al., 2003). This method requires the availability of segmentations of the structure(s) to be considered. Although overlap-based evaluation is intuitively reasonable it should be observed that it is limited by the quality of segmentations available and the type of structures that they represent. Large coarse structures (such as lung volumes) are typically well-captured by segmentations but rich details (such as vessel-trees within the lungs) are difficult to segment accurately and are therefore usually overlooked in evaluations which use overlap measures.

Other authors have measured registration accuracy based on point sets such as nodule positions (Blaffert and Wiemker, 2004; Betke et al., 2003), or manually annotated contours or landmarks (Vik et al., 2008; Castillo et al., 2009; Grachev et al., 1999; Crum et al., 2004; Wu et al., 2008; Boldea et al., 2008; Heath et al., 2007; Pevsner et al., 2006). Manual annotations are frequently small sparse point sets with poor distribution throughout the image. Larger sets of landmarks have been used by some authors, most notably Castillo et al. (2009) but these are typically required to be carried out by an expert observer which is expensive and impractical for a large set of images.

In (Jannin et al., 2002), Jannin et al. recommend the development of standardised validation procedures for medical image processing techniques including registration. In particular, validation using a common, publicly available set of validation data with corresponding ground-truth is advised. Unfortunately few such datasets are currently available due to the logistical difficulties of creating a comprehensive reference standard for registration. The 'Vanderbilt Dataset' (West et al., 1997) is a set of volumetric brain images available online as part of the Retrospective Image Registration Evaluation (RIRE) project. The reference standard for registration of these images is based on skull-implanted markers. A further set of 16 brain MR images is available from the Non-Rigid Image Registration Evaluation Project (NIREP) (Christensen et al., 2006) along with segmentation information. A single 4D lung CT dataset consisting of 10 3D images is supplied by the POPI-model (Vandemeulebroucke et al., 2007) including 41 landmarks identified in each 3D image. The University of Texas M. D. Anderson Cancer Center supplies 2 phases from each of 10 4D datasets, with 300 landmarks per phase (Castillo et al., 2009). For 2D–3D registration, data has been published by van de Kraats et al. (2005) and Tomazevic et al. (2004). We are not aware of any other freely available reference standards for registration.

One of the major obstacles to any group creating a large set of reference standard data, much less making it publicly available, is the amount of manual work involved in annotating the data. In this article a method is presented to formulate a registration reference standard in an efficient semi-automatic manner resulting in a well-distributed set of corresponding landmarks. The technique has been developed and demonstrated on pairs of temporal thoracic CT scans, however, in principle, it has the potential to be utilized in many other applications, particularly intra-subject, single modality registration problems. For inter-subject or multi-modality problems the same principle may be applied but modifications to certain parts of the algorithm would be required.

This method is designed specifically to overcome the problems of evaluating non-rigid registration techniques, although it may be equally well be applied to rigid or affine registration problems. The manual component of the reference standard construction could be completed by non-expert observers for our experiments, making it feasible to annotate large datasets without excessive consumption of expert resources. The ability to define correspondence in this way overcomes many of the difficulties usually involved in producing large sets of registration reference standard data. The software described in this article will be made publicly available on the website http://isiMatch.isi.uu.nl.

Reference standards have been constructed on 47 pairs of temporal thoracic CT scans, three pairs of brain MR scans and five pairs of CT scans with synthetic deformations. The opinions of the non-expert observers employed have been compared to those of a radiology expert in 5 cases. For the purposes of demonstrating the utility of the constructed reference standard data, it has been used in the evaluation of various registration procedures with different settings. Those settings where the registration performance is expected to be weak show decidedly worse results based on our evaluation. For other more successful registration algorithms, the ability to quantitatively analyse their results allows for the detection of subtle differences between them which might otherwise have been overlooked.

2. Materials and study setup

The principal set of data used in this work consists of a set of low-dose thoracic CT scans which form part of a lung cancer screening trial (Xu et al., 2006). Forty-seven subjects (44 male, 3 female, ages 51–74 years), each with a baseline and a follow-up scan (3–15 months apart) were chosen randomly from the screening trial database. All scans were obtained at full inspiration and without contrast injection on a 16 detector-row scanner (Mx8000 IDT or Brilliance 16P, Philips Medical Systems). Exposure settings were 30 mAs at 120 kVp for subjects weighing below 80 kg or 30 mAs at 140 kVp for those weighing over 80 kg. A soft reconstruction filter (Philips "B") was used. The scans have a per-slice resolution of 512×512, with the number of slices per scan varying from 374 to 579 (on average 462). Slice thickness is 1 mm with slice-spacing of 0.7 mm. Pixel spacing in the X and Y directions varies from 0.61 mm to 0.89 mm with an average spacing of 0.73 mm.

An additional set of MRI brain data from three subjects is also included. This data is taken from the SMART-MR study (Simons et al., 1999). The MR scans were made using a 1.5-T whole-body scanner (Gyroscan ACS-NT, Philips Medical Systems, Best, The Netherlands). The protocol consisted of a transversal T1-weighted gradient-echo sequence (repetition time (TR)/echo time (TE): 235/2 ms, flip angle 80). The image matrix size is 256×256 with 38 slices per scan. Slice thickness is 4.0 mm with an in-plane voxel size of 0.89 mm $\times$ 0.89 mm.

In order to establish a reference standard for the data described above, two medical students were employed to carry out the manual component of the ground-truth construction. Each student processed all scan pairs independently in order that the interobserver differences could be analysed.

3. Methods

In this section the methods to identify and match landmark locations in a pair of scans are described. Our main application is thoracic CT data and therefore the descriptions refer to the method

as it applies to this task. However the method is also later tested on brain MRI data as described in Section 4.4.

First a set of n well-distributed distinctive landmark locations are detected automatically in the baseline scan. The first s points (which are themselves also well distributed) are matched manually in the follow-up scan by an observer using a custom-made graphical user interface. The point pairs matched by the observer are used by the system to model the relationship between the baseline and follow-up scans. When s pairs have been completed the remaining $n - s$ matches are made fully automatically using the relationship model and a local block-matching refinement scheme.

Variable parameters mentioned in the text are listed in Table 1 along with the values assigned to them in the thoracic CT experiments and in the brain MR experiments described in Section 4.4. All parameter values were chosen empirically based on experimentation.

3.1. Automatic landmark detection

The initial step in setting the reference standard is to automatically determine a number of landmark locations in the baseline scan for each subject. It should be noted here that the landmark detection step will perform well only on data with sufficient structural detail such that corresponding points can be visually identified. Wörz and Rohr (2006) and Frantz et al. (2005) have published methods for determining significant anatomical landmarks based on an initial region of interest and on local image features. However, for our experiments, one of our main requirements was that the landmarks would be well-distributed throughout the region of interest (the lung volume in CT, and the brain tissue in MR). The nature of pulmonary anatomy, for example, is that most significant anatomical features are located around the mediastinal area, with very few if any points of interest in the outer regions of the lung close to the pleura. For this reason we developed a method of landmark detection to identify points which may not be structurally significant, but are sufficiently contrasted with their surroundings to allow an observer to visually identify the corresponding location on the follow-up scan.

The algorithm to detect landmarks described here is partially based on the work of Likar and Pernuš (1999). A lung mask is used to ensure that the points are located within the lungs since our application is concerned with registration of the lung volume only. (A mask is not required if the registration being evaluated is intended to register all visible structures in the image.) This mask was created by means of an automatic lung segmentation procedure described by Sluimer et al. (2005) and originally based on the work of Hu et al. (2001).

The algorithm to find landmarks automatically proceeds as follows: Points outside the lung volume are excluded from consideration. Within the lung volume, only every ith point in each direction is considered in order to improve computational efficiency. Points within d_p voxels of the pleural surface are also excluded since it is difficult to match these reliably in the follow-up scan due to the lack of local structure.

For all remaining points p at voxel location (x,y,z), with intensity $I(x,y,z)$, a distinctiveness value $D(p)$ estimating the dissimilarity of p with its surrounding region is calculated as follows:

1. An estimate of the gradient magnitude $G(p)$ at p is calculated by
$$G(p) = \sqrt{\boldsymbol{G_x}(\boldsymbol{p})^2 + \boldsymbol{G_y}(\boldsymbol{p})^2 + \boldsymbol{G_z}(\boldsymbol{p})^2},$$
where $\boldsymbol{G_x}(\boldsymbol{p})$, $\boldsymbol{G_y}(\boldsymbol{p})$ and $\boldsymbol{G_z}(\boldsymbol{p})$ are local directional gradients based on finite differences. $\boldsymbol{G_x}(\boldsymbol{p})$ is defined by
$$\boldsymbol{G_x}(\boldsymbol{p}) = \frac{I(x-1,y,z) - I(x+1,y,z)}{2},$$
and analogously for $\boldsymbol{G_y}(\boldsymbol{p})$ and $\boldsymbol{G_z}(\boldsymbol{p})$.
2. Points where $G(p)$ is below a threshold T_G are excluded from further processing as they are likely to be difficult to match reliably in the follow-up image.
3. Around each point p a hypothetical spherical surface with a radius of r_1 voxels is constructed (see Fig. 1) and m points, $q_1,\ldots,q_m$, uniformly distributed on the surface are selected using the technique of Saff and Kuijlaars (1997). A region of interest $ROI(q_i)$ around each point q_i is compared with the corresponding region of interest $ROI(p)$ around the original point p. The region of interest $ROI(a)$ of any point a is defined as a spherical kernel centred at a with a radius of r_2 voxels. The difference $\mathrm{Diff}(ROI(p),ROI(q_i))$ is defined as the average absolute difference of the corresponding voxel intensities in the two regions:
$$\mathrm{Diff}(ROI(p), ROI(q_i)) = \frac{1}{N}\sum_{k=1}^{N} |ROI(p)_k - ROI(q_i)_k|,$$
where $ROI(p)_k$ is the kth voxel in $ROI(p)$ and N is the number of voxels in $ROI(p)$ and in $ROI(q_i)$. Note that the values r_1 and r_2 should be selected based on an approximation of the sizes of structures and the distances between them in the particular image being processed.

Table 1
System parameters and the values used for the CT and MRI experiments described in this article. Distances are measured in voxels for the CT experiments and in millimetres for the MRI experiments as explained in Section 4.4.

Name	Description	CT	MRI
i	Only every ith point in each direction is considered when calculating landmark locations	5	2
d_p	Points within d_p of the mask boundary are not considered when calculating landmark locations	5	1
T_G	Gradient threshold below which points are not considered when calculating landmark locations	300	5
r_1	Radius of sphere constructed around each point under consideration when calculating landmark locations	8	8
m	The number of uniformly distributed points examined on the spherical surface when calculating landmark locations	45	45
r_2	The radius of the ROI around each of the m uniformly distributed points when calculating landmark locations	5	5
n	The number of landmark points to be selected from the candidate list	100	100
d_m	The initial minimum distance requirement used when ordering the landmark points such that they are well-spaced	400	400
c_1	The length of the cube side for the cubic region defining candidate points to be considered during block-matching	13	13
c_2	The length of the cube side for the cubic region compared during block-matching	13	13
T_{SSD}	The threshold used in block matching for the root mean squared difference per voxel over the block being matched	165	165
s	The number of points required to be manually matched	30	30
x	Threshold used in establishing the accuracy of the trained system. x of the y most recently matched points are checked to compare system guess and observer chosen match	9	9
y	Threshold used in establishing the accuracy of the trained system. x of the y most recently matched points are checked to compare system guess and observer chosen match	10	10
d_a	The threshold distance for deciding accuracy in comparing system guess locations with observer chosen matches	$\sqrt{6}$	$\sqrt{18}$

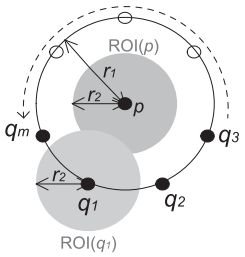

Fig. 1. Comparing point p with its surroundings as part of the process to measure distinctiveness. A spherical surface around p is constructed and m points q_i are identified on it. ROI(p) is compared with each ROI(q_i).

4. The distinctiveness value $D(p)$ is calculated for each point p as follows:

$$D(p) = \frac{G(p)}{max_j(G(p_j))} \frac{1}{m} \sum_{i=1}^{m} \mathrm{Diff}(ROI(p), ROI(q_i)),$$

where j is the total number of points for which we calculate $D(p)$ in this scan.

A large number of points in each baseline scan are labelled with a distinctiveness value in this way. A final selection of n landmarks will be chosen from these, based not only on their distinctiveness values but also on their locations. An even distribution of the landmarks throughout the lungs is required and furthermore, since the points will later be used in the creation of a thin-plate-spline (see Section 3.2) the ordering of the chosen points should be such that each one is as far away as possible from preceding selected points.

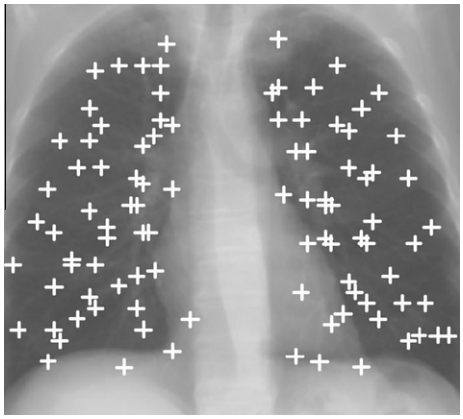

Fig. 2. A projection view of all landmarks identified in a scan. Marker sizes have been increased for visualisation.

The list of processed points p is initially ordered with the most distinctive points first and n landmarks are chosen, with a well-distributed ordering, as follows:

1. The most distinctive point available is selected as a landmark as long as it is at least d_m voxels in distance from every other point selected so far.
2. If the end of the list is reached then no more points meet this requirement. Set $d_m = d_m - 10$ (voxels) and repeat step 1.
3. Continue until n landmarks have been selected.

A projection view of all the landmarks selected for a scan is shown in Fig. 2 while Fig. 3 shows some examples of landmark locations.

3.2. Establishing landmark correspondence

A semi-automatic system was developed to accurately match the voxels identified as landmarks in the baseline scan with voxels at the corresponding anatomic locations in the follow-up scan. The observers were firstly required to match a subset of the landmarks manually using a custom-made graphical interface. Point pairs selected manually were used to create a thin-plate-spline (TPS) model of the deformation between the two scans in question. Other models such as the elastic body spline (Davis et al., 1997; Kohlrausch et al., 2005; Wörz and Rohr, 2006) might also have been substituted at this point. We chose to use the TPS model in order that only the user-selected points would influence the model, and no other parameters such as tissue properties could alter the outcome. This decision was based on the assumption that the TPS model would be sufficient to describe the deformation between the images, particularly when combined with the subsequent block-matching refinement step.

After s points had been manually matched and provided that the TPS model was deemed sufficiently accurate, the system matched the remaining points automatically. These steps are described in more detail in the remainder of this section. The entire annotation procedure took 20–30 min per scan pair for the thoracic CT data.

3.2.1. Graphical user interface

The graphical user interface was designed to allow the observer to view the landmark α in question on the baseline scan in all three orthogonal directions simultaneously. The location of the landmark α is identified by a red crosshair symbol in each of the three images. These three images are located on the upper half of the screen while on the lower half the follow-up scan is presented in a similar fashion with three orthogonal views visible. No identifying crosshair is initially shown in these images although the

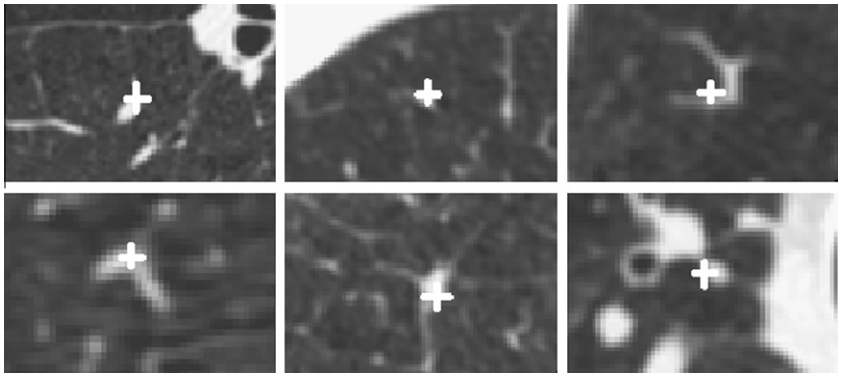

Fig. 3. Six sample landmark locations viewed close-up in axial direction. Marker sizes have been increased for visualisation.

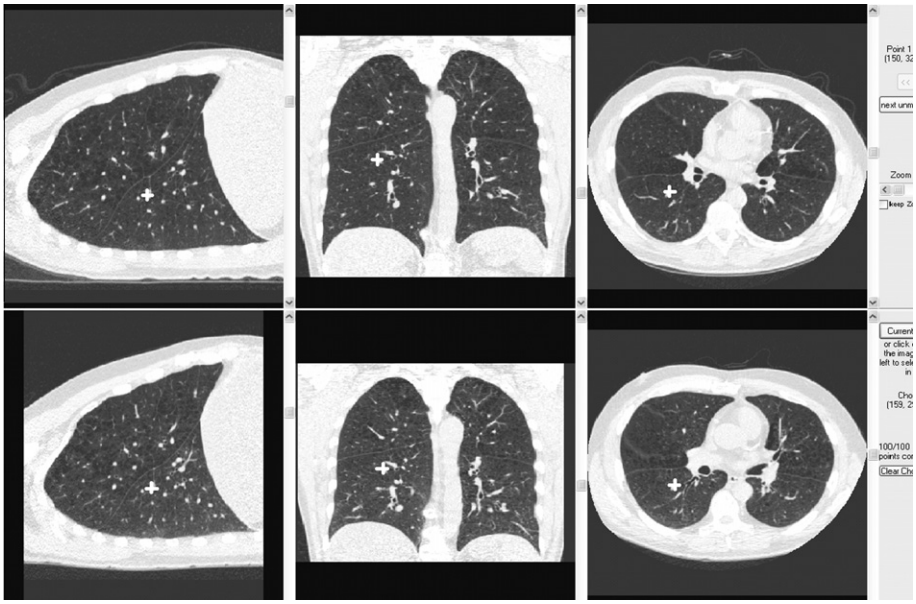

Fig. 4. The graphical user interface used to match points in a baseline scan (top row images) and a follow-up scan (bottom row images). The images are zoomed out as when the observer first begins with point-matching. Marker sizes have been increased for visualisation.

system attempts to present the most likely slices for the matching location. Further information on the determination of the most likely matching location is given in the next section. Screen-shots of the system are shown in Figs. 4 and 5. The user is allowed to manually select the matching landmark location β_{man} in the follow-up scan in two ways:

1. By clicking on any point in one of the three orthogonal views of the follow-up scan the 3D location of the point β_{man} is selected.
2. By scrolling through the three orthogonal views of the follow-up scan individually, the most appropriate slice in each direction is identified. When satisfied the observer clicks a button to select the point identified by the three visible slices.

After selection of the matching location β_{man} a red crosshair icon is placed in the appropriate position in the follow-up image to indicate the observer's choice. The observers were encouraged to view the locations α and β_{man} at various zoom-levels and to confirm their final choice at the highest possible zoom-level where individual voxels were clearly visible. They were permitted to repeatedly re-locate their matched landmark until they were satisfied with their choice. In cases where the observer was unable to find a satisfactory match they were instructed to place the match in the best location they could identify and to check a box to indicate their uncertainty. Those points where the observer was uncertain of the match location were not included in the TPS model in order to retain its integrity. The

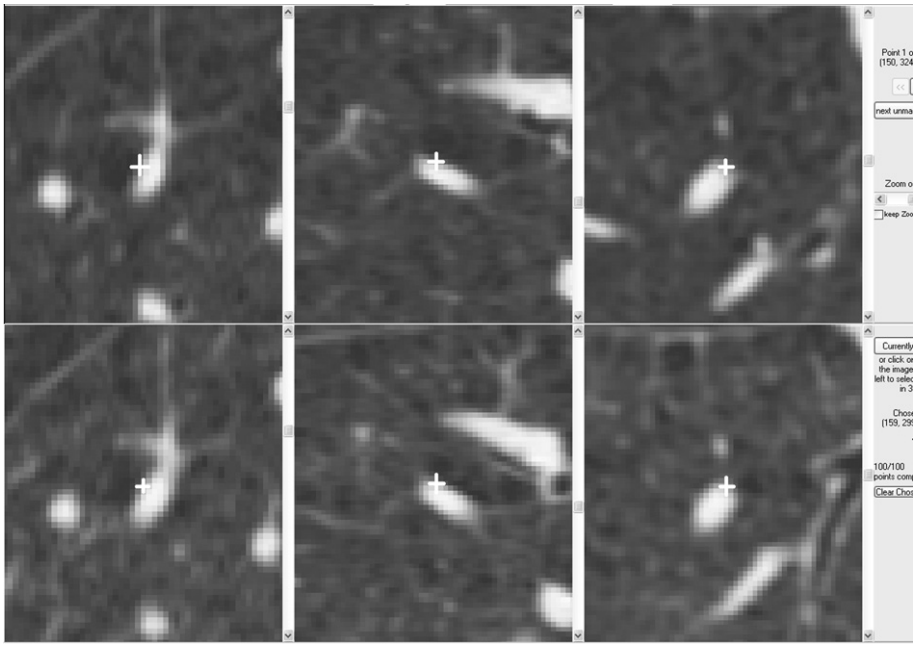

Fig. 5. The graphical user interface used to match points in a baseline scan (top row images) and a follow-up scan (bottom row images). The images are shown zoomed in on a particular point. Marker sizes have been increased for visualisation.

uncertain point pairs are otherwise treated as standard throughout the procedure.

3.2.2. Automatic landmark matching

The matching pairs of landmark correspondences manually annotated by the observer are used in the formation of a thin-plate-spline (Bookstein, 1989) (TPS) warping of the follow-up image. This warping process begins as soon as the observer has matched the first four point pairs. It must be stressed that a warped image is not displayed to the user at any time. The TPS is used only internally to represent the relationship between the baseline and follow-up images, and to help with predicting future match locations. Each point pair (α, β_{man}) manually annotated by the observer is added to the TPS unless the observer indicates uncertainty. The accuracy of the TPS is thereby progressively improved. When a new landmark point is presented to the observer for manual matching the system makes an estimate β_{est} of where the anatomic match will be located in the follow-up scan as follows:

1. The TPS warping is interpolated to get an initial estimate $\beta_{est_{init}}$ of the location in the follow-up scan corresponding to the landmark α.
2. A local block-matching search to improve upon this initial estimate is performed in the region around $\beta_{est_{init}}$. This scheme is similar to that described in Wiemker et al. (2008) and proceeds as follows (see Fig. 6):
   (a) All voxels β_{est_k} in a cube of side c_1 voxels around $\beta_{est_{init}}$ are considered as candidates.
   (b) Cubic regions of interest ROI(α) and ROI(β_{est_k}) with sides of length c_2 voxels are defined around the landmark point α and the point β_{est_k} under investigation.
   (c) The β_{est_k} where the sum of squared differences (SSD) between intensities in the regions of interest, SSD(ROI(α), ROI(β_{est_k})) is minimal is selected as the final system estimate β_{est}.

The location β_{est} of the estimated match is used to determine which slices from the follow-up scan should be displayed initially to the observer. Therefore, as the TPS warping becomes more accurate, the task of the observer becomes easier, with the initially displayed slices providing increasingly accurate starting points.

After some time the TPS warping and block-matching scheme is sufficiently accurate to enable the system to proceed with matching the remaining landmarks without user interaction. Automatic matching is permitted when

1. The observer has manually matched at least s landmarks α (including those where the match was uncertain) with corresponding locations β_{man}, and
2. The system has estimated x of the previous y matches such that the distance between the estimate β_{est} and the location β_{man} indicated by the observer was less than or equal to d_a voxels.[1]

When these conditions are satisfied a button appears on the screen which the observer clicks to match all remaining points automatically. The automatic matching searches for a match β_{auto} using the TPS warp and block-matching scheme exactly as described above in the search for β_{est}. Note that points found automatically in this way are not added to the TPS and hence do not influence the locations of further automatically matched points. In cases where the block-matching finds that the root mean squared difference per voxel over the block being matched exceeds a threshold (T_{SSD}) then the match β_{auto} is considered uncertain and the landmark α is returned unmatched to the observer. Such points must be matched manually by the observer, however this occurred rarely in our experiments (see Fig. 7) and the system estimate in those cases was often correct allowing the manual match to be made without difficulty.

The system required a match to be made for every landmark point, although it may sometimes be the case that a true match does not exist (for example in the case where the landmark is on an artifact or a growth which is only present in only one image). This is a relatively rare occurrence in most datasets and there is no perfect way to deal with the issue. In this case we chose to force the user to match the point, since a registration algorithm will be similarly forced to specify a correspondence at every location. It is, however, equally acceptable to instruct the user not to select any corresponding point if there are genuine anatomical differences.

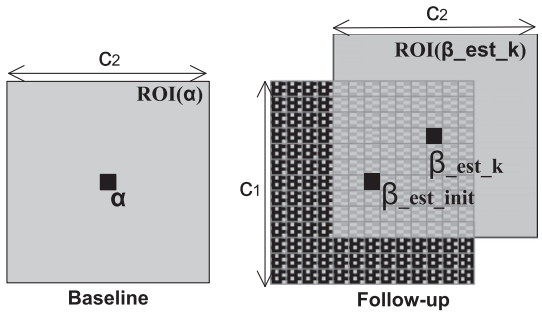

Fig. 6. The block-matching concept (with squares used to represent the cubes). The voxels indicated with patterned texture are those which will be considered as candidate voxels (β_{est_k}). For each of these locations the region $ROI(\beta_{est_k})$ will be compared to the region $ROI(\alpha)$ and the location where the SSD of intensities is minimum will be selected.

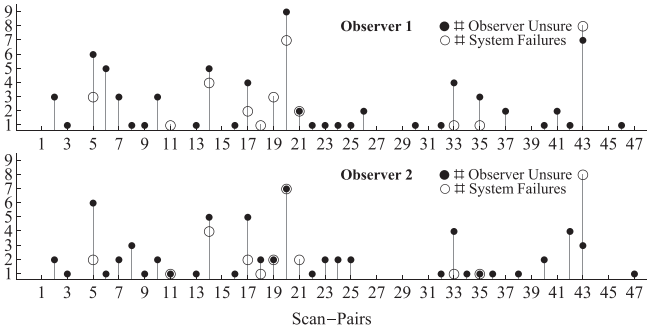

Fig. 7. Statistics for each observer. The number of matches marked 'unsure' and the number of automatic match failures (T_{SSD} exceeded) per scan pair for Observer 1 (top) and Observer 2 (bottom).

3.3. Registration methods

In order to demonstrate the use of the reference standard data a number of registrations were carried out and then analysed as described in Section 4.5. All registrations in this study were carried out using `elastix` version 3.9[2] which is a registration toolkit based on the National Library of Medicine Insight Segmentation and Registration Toolkit (ITK). Although a single registration

[1] Note that d_a may reasonably be set at a value greater than 0 for most tasks. In practice the choice between a particular voxel and its close neighbours is frequently difficult for an observer and the final decision may be somewhat arbitrary. The system may therefore be deemed to be correct if it chooses a location very close to the observer annotation.

[2] http://elastix.isi.uu.nl.

package is used, the elastix toolkit provides numerous options for the registration procedure including several similarity measures, image mask support and both affine and non-rigid registration components. This allows us to specify six different registration configurations with varying results and demonstrate their evaluation. The non-rigid transformations are modelled by a B-Spline grid (Unser, 1999; Rueckert et al., 1999).

In the experiments for this study we exploit our ability to distinguish quantitatively between registration methods using the reference standard data. This is achieved by registering the data several times with various elastix configurations and comparing the outcomes with the reference standard. The basic registration settings from which all subsequent variations arose are described below.

Prior to registration the baseline and follow-up scans were down-sampled in order to improve speed and reduce memory consumption. The down-sampling was by means of block-averaging such that the matrix size of 512×512 in the original images was reduced to 256×256, with the number of slices reduced to form isotropically sampled data. The down-sampled follow-up scan (source image) was registered to the down-sampled baseline scan (target image) and the resulting transformation was subsequently applied to the full resolution follow-up scan. Each registration consisted of an initial affine registration step followed by a non-rigid registration to model the elastic behaviour of the lung tissue. Both steps involved a multi-resolution strategy with four resolution levels for the affine procedure and five for the non-rigid procedure. A mutual information cost function (Thévenaz and Unser, 2000) was used in both cases along with a stochastic gradient descent optimizer (Klein et al., 2007). Termination of the optimization procedure occurred in each resolution after a fixed number of iterations, set at 512. The grid-size varied per resolution level with the finest grid at the last level having a spacing of 5 mm in each dimension. Lung mask images were used to ensure that only anatomy within the lungs was registered, thus excluding the ribs, heart and other confounding structures. These masks were created by the same automatic technique mentioned in Section 3.1 (Sluimer et al., 2005).

All registration experiments carried out are listed in Table 2 with an explanation of any changes to the basic settings described above. The basic setting (BS) had been experimentally determined to be relatively fast and accurate on this type of data in the past. The choice of inclusion of other configurations is discussed in Section 5.

Table 2
Registration experiments.

Registration	ID	Explanation
Basic	BS	Settings as described in Section 3.3
Affine-only	AF	No non-rigid registration is carried out
No-masks	NM	No image masks are used to specify the regions of the image to be registered
Mean-squares	MS	Mean sum of squared differences is used as a similarity measure in place of mutual information
Cross-correlation	CC	Normalised cross-correlation is used as a similarity measure in place of mutual information
Full resolution	FR	Original images are not down-sampled prior to registration

4. Experiments and results

In this section a variety of experiments are described. In Section 4.1 the process of constructing the reference standard on thoracic CT data is analysed, describing the accuracy of the system and the interobserver differences. Section 4.2 outlines an experiment whereby a radiology expert was asked to annotate five scan pairs and the results are compared with those of medical student observers. In Section 4.3 a process for synthetically warping a scan is described and observer results based on the synthetically warped data (where the real ground truth is known) are shown. Section 4.4 describes experiments on brain MR data, and Section 4.5 analyses the results of various registration procedures using the reference standard defined on the 47 thoracic CT pairs.

4.1. Reference standard construction results

4.1.1. System and observer behaviour

As described in Section 3.2 there are two criteria which must be met before the system can begin automatic matching: firstly the observer must have made at least s matches fully manually, and secondly the system guess must have been demonstrated to be accurate in x of the previous y points matched. In this study s was set at 30 and 47 scan pairs were included. When the required 30 manual matches had been made the second criterion of system accuracy was usually also satisfied. An exception to this occurred just once for each observer, forcing the observer to make additional manual matches in order to improve the system accuracy. (Observer 1: scan pair 5, 38 manual matches, Observer 2: scan pair 43, 31 manual matches). This gives an initial indication that the system is usually well trained after 30 points have been manually matched.

Fig. 7 illustrates the number of matches marked 'unsure' and the number of times the automatic system failed to find a reliable matching point (T_{SSD} exceeded) for each of the 47 scan pairs processed. It can be seen that in general the observers found similar levels of difficulty in each scan pair, and that in the small number of more difficult scan pairs (where the observer is often unsure) the automatic system also has an increased failure rate.

It is important to assess the accuracy of the system guess β_{est} and the rate at which the system improves in its ability to predict match locations. The system guesses were stored for this purpose and later compared to the manual annotations β_{man} made by the observers. The average Euclidean distance δ_j between the manually annotated location β_{man_j} and the system guess β_{est_j} after j points have been manually matched was calculated over all 47 scan pairs. The values

$$\delta_j = \frac{1}{47}\sum_{i=1}^{47}[\beta_{est_j} - \beta_{man_j}]_{scanPair_i}$$

were calculated for j from 1 to 30. The results of this analysis are depicted in Fig. 8, where δ_j is plotted against j for each observer to illustrate the system 'learning curve' as increasing numbers of point pairs are annotated. Fig. 9 breaks down the average value δ_j to show all distances $\beta_{est_j} - \beta_{man_j}$ in box-whisker plot form.

4.1.2. Interobserver differences

The interobserver differences were analysed to verify the ability of observers and of the system to find reproducible corresponding anatomic locations for the landmarks. The landmarks were presented to the observers in the same order, therefore in general the same points are matched manually by both observers. However, a small number of points (only 11 points in all 47 scans in this case) have two different 'match-types' i.e. they are marked manually by one observer but not by the other. There were two possible ways for this to occur: (1) if one observer was required to match more points manually (before automatic matching could proceed) than the other observer, or (2) if automatic point matching exceeded the threshold T_{SSD} on a particular point for one observer only, thus requiring him to make that match manually. In Fig. 10

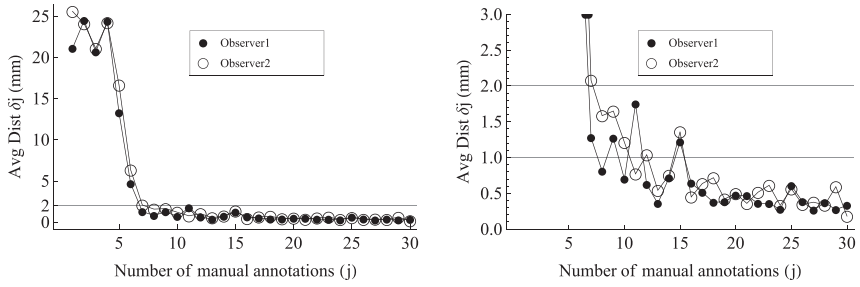

Fig. 8. The system 'learning curve'-values of δ_j plotted against j (left). The same graph focusing more closely on the end region (right).

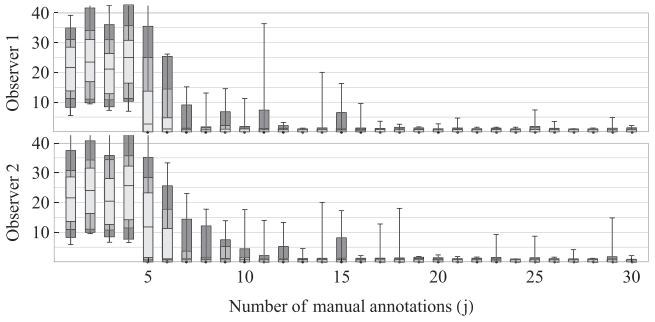

Fig. 9. Box-whisker plots showing the distances $\beta_{est_j} - \beta_{man_j}$ for all scan pairs plotted against number of manual annotations. Top: Observer 1, bottom: Observer 2. The lightest grey colour indicates the boundaries of the central 50% of the data, mid-grey the central 75% and the darkest grey the central 90%. The horizontal line within the central 50% marks the median value. Outliers are included in the whisker length.

the interobserver differences in mm for all 4700 landmarks are illustrated, categorised by match-type. As expected, points which were marked automatically in both cases are considerably more likely to have differences of 0 mm than those which were marked manually, since in the automatic case a local search for the lowest *SSD* is performed. For the registration analysis described in Section 4.5 the (339) points where the interobserver difference was greater than 1 mm are disregarded because of the uncertainty of the reference standard in these cases.

4.2. Expert observer annotations

Since observers 1 and 2 were medical students with no specific training in reading thoracic CT data an expert observer was asked to annotate the first five scan pairs in order to compare his results with those of the untrained observers. The expert in this case is a radiologist in training (a physician with 3 years of experience in radiology, particularly in lung CT evaluation). The expert observer annotated the five scan pairs independently in exactly the same way that observers 1 and 2 had done. The interobserver differences for points manually annotated by all three observers (1, 2 and expert) are shown in Fig. 11.

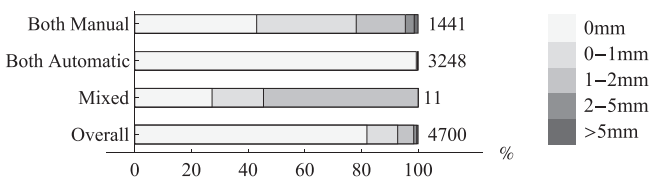

Fig. 10. Interobserver differences categorised by match-types. The number at the end of each bar signifies the total number of points in the category.

4.3. Synthetic warping

In this experiment the performance of the observers is assessed in a situation where the deformation between the images in question is synthetic and the real ground-truth is therefore available for comparison. At the time of this experiment observers 1 and 2 were no longer available and were replaced by two other medical students who had also had some experience with annotating lung data using the system described in this work. They will hereafter be referred to as observers 3 and 4.

A synthetic warp was performed on the baseline scan $Scan_B$ of each of scan pairs 1–5. In each case the warp was modelled by a thin-plate-spline (TPS). The point pairs used to create the TPS model were those pairs which were manually annotated by the expert observer as described in Section 4.2. Points which were marked unsure or were automatically matched were excluded. Using this TPS model, a synthetically warped version $Scan_{BW}$ of $Scan_B$ was produced. ($Scan_{BW}$ bore some resemblance to the follow-up scan $Scan_F$ from the original scan pair as a result of the warping which was used to create it, however they were not identical.) The outer regions of $Scan_{BW}$ were cropped to exclude locations which the warping process had been unable to fill with data values and the scan was inspected to ensure that it appeared realistic. A coronal slice from the baseline scan $Scan_B$ of scan pair 1, and the associated warped scan $Scan_{BW}$ are shown in Fig. 12.

The pair $Scan_B$ and $Scan_{BW}$ were presented to observers 3 and 4 without informing them that one of the scans was synthetic. They were asked to match the landmark points as normal. Their matching locations were compared with the known ground-truth given by the expert observer matching points which were used to create the synthetic warp. The distances between the observer matching points and the ground-truth are shown in Fig. 13.

4.4. Brain MRI data

The system described in this article was developed specifically for a thoracic CT application. In order to test its performance on an alternative type of data, three sets of brain MRI data as described in Section 2 were obtained with a baseline and follow-up scan for each patient.

Some modifications to the system were required due to the anisotropic nature of the MRI data used (voxel sizes $0.89 \times 0.89 \times 4.0$ mm). Firstly all distances which had previously been measured in voxels on the almost isotropic lung data were now required to be measured in millimetres. Secondly the finite difference method for estimating gradient magnitude was altered to include weights according to the voxel sizes in each dimension. The new gradient magnitude estimate is calculated by

$$G(p) = \sqrt{\boldsymbol{G_x}(\boldsymbol{p})^2 + \boldsymbol{G_y}(\boldsymbol{p})^2 + \boldsymbol{G_z}(\boldsymbol{p})^2},$$

as previously, but in this case $\boldsymbol{G_x}(\boldsymbol{p})$ is defined by

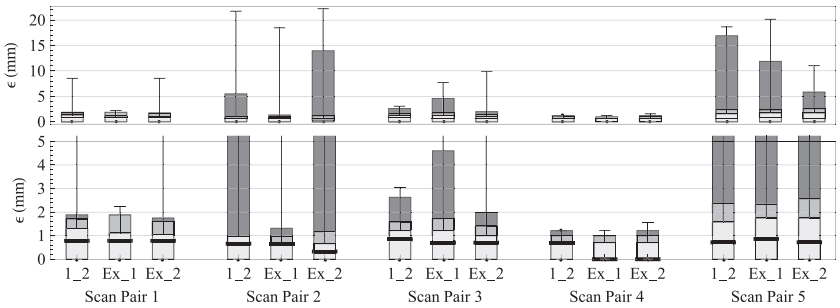

Fig. 11. Interobserver distances for manually matched points for the first five scan pairs shown in full (above) and in closer view (below). Within each scan-pair grouping the three columns from left to right represent Observer 1/Observer 2 distances, Expert/Observer 1 distances and Expert/Observer 2 distances respectively. The lightest grey colour indicates the boundaries of the central 50% of the data, mid-grey the central 75% and the darkest grey the central 90%. The horizontal line within the central 50% marks the median value. Outliers are included in the whisker length.

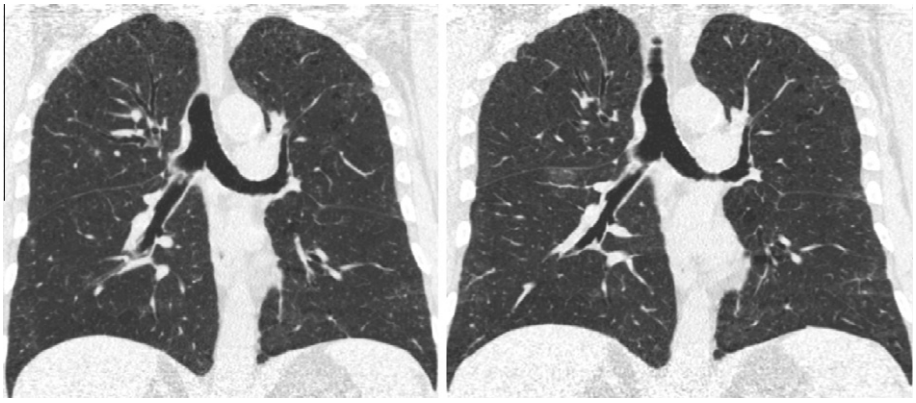

Fig. 12. Left: a coronal slice from the baseline scan in scan pair 1. Right: a coronal slice from the warped version of the same scan created using the expert observer annotations.

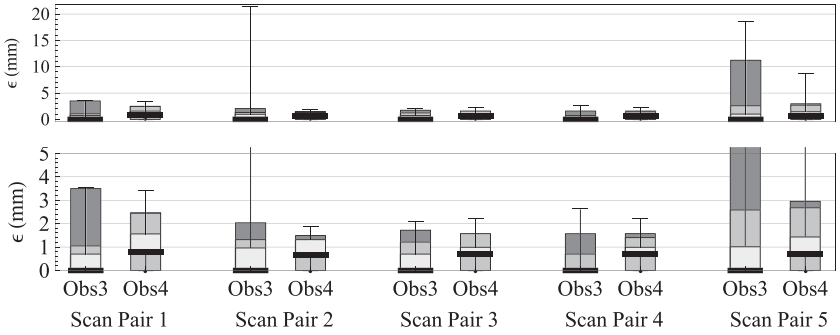

Fig. 13. Observer-truth distances for manually matched points on the synthetically warped datasets. Distance box plots are shown in full (above) and in closer view (below). Within each scan-pair grouping the left column represents distances for observer 3 and the right column for observer 4. The lightest grey colour indicates the boundaries of the central 50% of the data, mid-grey the central 75% and the darkest grey the central 90%. The horizontal line within the central 50% marks the median value. Outliers are included in the whisker length.

$$\boldsymbol{G_x}(\boldsymbol{p}) = W_x\left(\frac{I(x-1,y,z) - I(x+1,y,z)}{2}\right),$$

where W_x is a weighting defined by the inverse of the distance between the voxels at $(x-1,y,z)$ and $(x+1,y,z)$.

$$W_x = \frac{1}{2v_x},$$

where v_x is the voxel size in the x direction. $\boldsymbol{G_y}(\boldsymbol{p})$ and $\boldsymbol{G_z}(\boldsymbol{p})$ are calculated analogously.

No automatic segmentation software was available for these images so masks were drawn by hand to denote in which regions distinctive points should be located. The masks were designed to include brain tissue but exclude the skull and cerebrospinal fluid. These are excluded for the same reason we excluded points close to the lung boundary, as point matching in those regions is unreliable. Note that an automatically generated mask of the entire patient anatomy would also have been suitable for use provided that the distance d_p from the mask boundary within which points should not be marked was set appropriately.

Landmark detection was carried out as described in Section 3.1 with the parameters set as shown in Table 1. Detected landmark points for one of the MRI scans are shown in Fig. 14.

MR images, unlike CT, do not have a fixed relationship between tissue-type and grey-value, as illustrated in Fig. 15. Block-matching

with SSD as a similarity measure would therefore be unreliable in MR data and the system was used with TPS warping only in these experiments.

Observers 3 and 4 annotated points in the three scan pairs exactly as before with parameters set as shown in Table 1. The minimum requirement to count a system guess as 'accurate' (d_a) was increased to allow for the fact that no block matching was used and guesses were therefore expected to be slightly less accurate.

The system learning curve based on the experiments using TPS only is shown in Fig. 19a and interobserver differences for the two observers are shown in Fig. 19b.

4.5. Registration performance analysis

In this section the reference standard data which was constructed for the main dataset of 47 thoracic CT scan pairs is used in the evaluation of the various registration procedures listed in Table 2. The aim of the evaluation in this case is not to determine which registration is superior (a question which is largely application dependent), but rather to illustrate the utility of the reference standard data in obtaining a quantitative assessment. The performances of the various registration procedures are discussed in Section 5.

The registration of an image-pair results in a transform T which maps from locations in the domain of the baseline scan (the target image) to locations in the domain of the follow-up scan (the source image). In order to judge the accuracy of a registration method, T is applied to each of the landmark points α defined in the baseline scan. For an accurate registration we expect $T(\alpha) \approx \beta$, where β is the matching point marked during reference standard formulation. For all points β_{obs1} marked (manually or automatically) by Observer 1 the registration error $\epsilon(T(\alpha), \beta_{obs1})$ is defined as the Euclidean distance between $T(\alpha)$ and β_{obs1}.

In Fig. 16 box plots are presented illustrating the registration-Observer 1 distances ϵ over all scan pairs for each registration procedure. The leftmost plot shows the interobserver differences for reference. Only those points where the interobserver difference was less than 1 mm are used as part of the reference standard.

The same registration error measurements are subdivided in Fig. 17 to depict the manual and automatic components of the reference standard separately. It appears that for the purposes of registration evaluation there is virtually no difference in the quality of reference standard points which were manually chosen and those which were selected by the system during the automatic phase.

The performance of the registration methods investigated was also considered on a per-scan basis. Registration error data from the first 25 scan pairs is shown in Fig. 18 for all registration

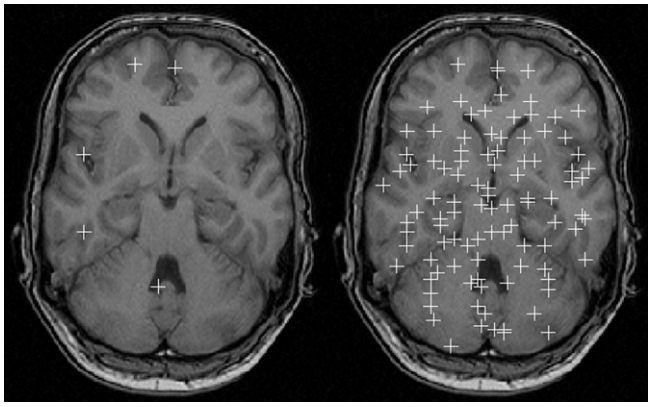

Fig. 14. Left: an axial slice from a brain MRI image showing the landmark points located in that slice (left) and all landmark points for the whole scan projected onto that slice (right).

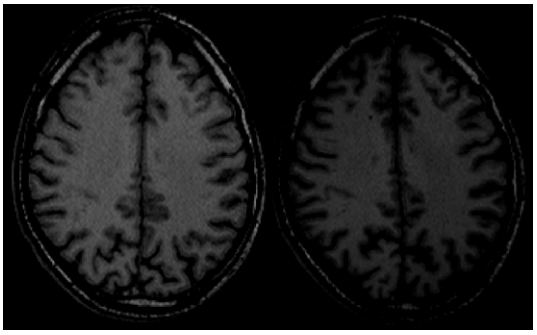

Fig. 15. The baseline (left) and follow-up (right) scans for pair 2 of the brain MR data. Contrast is set identically in both scans with a narrow window width to illustrate that the same tissues are represented by different grey-values in each scan.

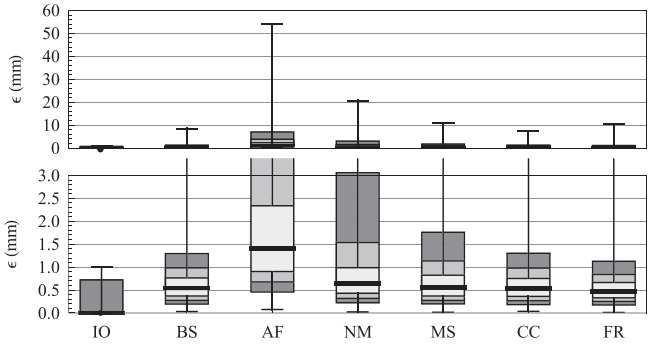

Fig. 17. Interobserver (IO) and registration-Observer 1 distances (ϵ) shown in full (above) and in closer view (below). Each box plot is limited to either manually matched points (M) or automatically matched points (A). The acronyms referring to registration methods are explained in Table 2. The lightest grey colour indicates the boundaries of the central 50% of the data, mid-grey the central 75% and the darkest grey the central 90%. The horizontal line within the central 50% marks the median value. Outliers are included in the whisker length.

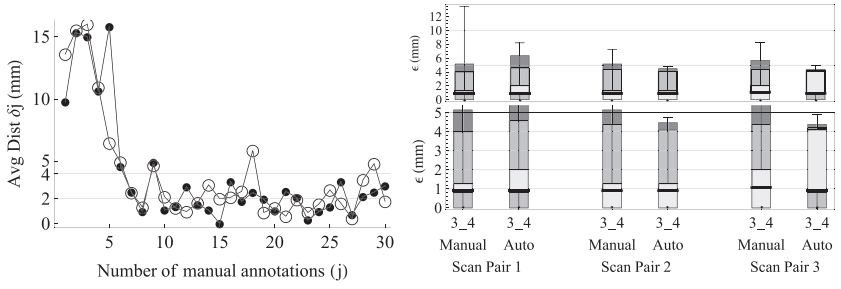

Fig. 16. Interobserver (IO) and registration-Observer 1 distances (ϵ) shown in full (above) and in closer view (below). The acronyms referring to registration methods are explained in Table 2. The lightest grey colour indicates the boundaries of the central 50% of the data, mid-grey the central 75% and the darkest grey the central 90%. The horizontal line within the central 50% marks the median value. Outliers are included in the whisker length.

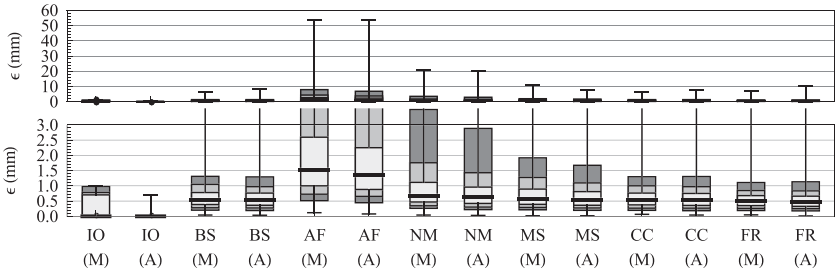

Fig. 18. Registration-Observer 1 distances (ϵ) shown individually for the first 25 scan pairs. Registration settings from top to bottom: basic, affine-only, no-masks, mean-squares, cross-correlation and full-resolution. The lightest grey colour indicates the boundaries of the central 50% of the data, mid-grey the central 75% and the darkest grey the central 90%. The horizontal line within the central 50% marks the median value. Outliers are included in the whisker length.

Table 3
Means and standard deviations of the times taken for registration.

Registration	Mean time (min)	Standard deviation
BS	4.20	0.26
AF	1.37	0.09
NM	4.11	0.91
MS	2.87	0.08
CC	3.04	0.09
FR	12.13	0.82

methods. Similar results are observed for the remaining 22 scan pairs although the data has been omitted for brevity.

Finally, when comparing registration algorithms, the time taken to perform the registration is often important in deciding the whether using the algorithm would be practicable, for example in a clinical situation. To give a guideline as to the computational cost of the various registration algorithms, Table 3 lists the mean and standard deviation of the time taken to complete registration for each of the methods listed in Table 2. All experiments were carried out on a desktop PC running Microsoft Windows Server 2003 with an Intel Core2 Quad processor (2.4 GHz) and 6640 MB of RAM.

5. Discussion

A semi-automatic system for reference standard formulation in registration has been presented. In the thoracic CT experiments described the system defines a well-distributed set of corresponding landmark points with limited interaction from non-expert observers. The accuracy of the defined correspondences is implied by the independent observations of two observers with 98.5% of interobserver differences below 2 mm and 92.8% within 1 mm (see Fig. 10, 'Overall'). For manually matched points the differences are slightly higher (78% within 1 mm) as would be expected. The finite resolution ($\approx$0.7 mm) of the image data makes it difficult for an observer to select a particular voxel (above all its neighbours) to be the correct matching point.

The ability of the system to model deformation and predict anatomic matches is demonstrated in Fig. 8 where the increasing accuracy of the system guesses is illustrated. It is clear that the ability of the system to predict corresponding point locations improves rapidly after the first five to six points have been annotated. The average error in the system guess after this stage remains below 2 mm for both observers. Similarly, after 15 points have been manually matched the system guess is on average always within 1 mm of the observer decision. The increasing accuracy of the system accelerates the manual phase of the matching procedure by providing the observer with ever more precise starting points and ultimately enables the introduction of fully automatic matching. The box plots shown in Fig. 9 demonstrate that in general the actual distance values deviate less from the median values as more points are manually annotated. In Fig. 17 the registration evaluations based on fully manual and fully automatic point pairs are shown to be almost identical, indicating that the automatically matched points are equally useful in the evaluation of registration.

In Section 4.2 the opinions of medical students are compared with the opinions of a radiology expert in order to ascertain whether the medical students possibly lacked the expertise to define point correspondences. Fig. 11 shows the results of this experiment. It can be seen that there is no case where the medical students (observers 1 and 2) were both at odds with the expert opinion while in agreement with each other. This implies that their agreement on points was not coincidental or due to their lack of expertise. In scan pair 2, Observer 2 is seen to disagree frequently with both Observer 1 and the expert, while there is excellent Observer 1/expert agreement. This implies that Observer 2 made a higher than normal number of errors in this scan pair. Scan pair 5 shows a relatively high level of interobserver differences between all three opinions, indicating that this scan pair was more difficult than the others. This is backed up by other results discussed later in this section. It is clear that interobserver differences may be high due to difficulties with a particular dataset or due to the lapse in concentration or attention to detail of an observer. Setting a study up initially with two observers is a reasonable choice in order to determine how much disagreement occurs. If discarding points with high interobserver differences is not desirable then the addition of extra observers to the study (possibly even for a limited number of datasets) is a good option. Except in the case of extremely difficult datasets which may never be reliably annotated this should be sufficient to establish a reliable reference standard.

Synthetic warpings were generated for five scans as described in Section 4.3 and used to compare the observer opinions with a known ground-truth. In Fig. 13 the distances between manual observer marks and the known ground-truth are shown. It can be seen that the median distance is below 1 mm for both observers in all five cases. 90% of the distances are within 2 mm for scan pairs 2, 3 and 4. Scan pairs 1 and 5 proved to be more difficult, with observer 3 in particular making more errors. However it is worth noting that we have included all manual points here, including those where the observers disagreed. If we restrict the analysis to points where the observers agree within 1 mm then the agreement with ground-truth is also significantly better. For scan pairs 2 and 5 the median distance drops to below 1mm while the maximum distance from the ground-truth is just 2 mm. This demonstrates the benefits of having two independent observer opinions available.

The system was also tested on brain MR data as described in Section 4.4. One issue encountered here was that block-matching using SSD was not useful since tissues were represented by different grey-values in different images. Therefore, for MR data, and other data where tissue value ranges may vary, it would be advisable to alter the block-matching scheme as appropriate for the data. For example, if it is known that there is a linear relationship

between the intensities in the images then a similarity measure such as normalized cross-correlation might be used in place of the SSD measure. The system behaviour was evaluated on the brain MR data without using any block-matching, and the results are shown in Fig. 19. The system 'learning curve' demonstrates that the TPS system became reasonably accurate after about six point pairs had been manually matched. Subsequently, the system guess was usually within 4 mm of the final selection made by the observer. If a block-matching refinement scheme was added to this system we would expect the guesses to become more accurate still. The interobserver distances shown on the right of Fig. 19 show that the median interobserver difference was about 1 mm in all cases, while 90% of the distances were within about 5 mm. Interobserver distances would be expected to be somewhat higher than those in the thoracic CT data since the slice thickness of 4 mm in the MR data caused a stronger partial volume effect. This made precise selection difficult for the observers in some cases. Overall, the performance on the brain MR data is good, with the observers reporting that the system guesses were very useful. Using a revised block-matching refinement scheme there is the potential for excellent results on this data.

In general the system has the potential to be used on many types of data, possibly with minor modifications. At a minimum the system parameters listed in Table 1 will need to be adjusted experimentally according to the type of data being used. As already demonstrated using brain MR the block-matching similarity measure of SSD will not be suitable for all data types. Scan pairs exhibiting more severe deformations (for example between full inspiration and full expiration in thoracic CT) are likely to be more difficult than those scans examined in this work, however we do not anticipate severe problems with increased deformations apart from a longer system training time. At present it seems that the TPS model is sufficient to describe the types of deformations that may be encountered in medical scans, however this assumption needs to be verified for each new type of data being processed. More challenging tasks such as inter-modality or inter-subject

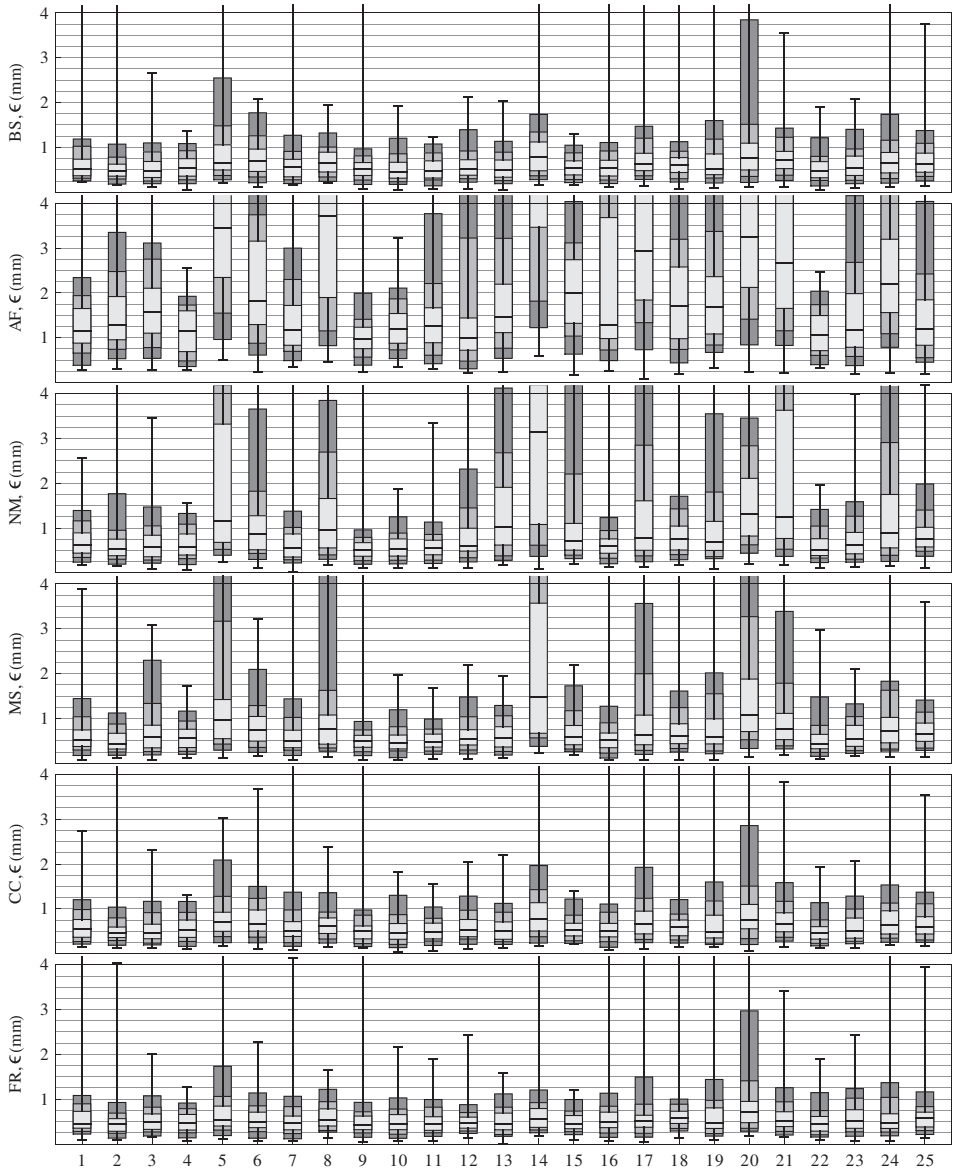

Fig. 19. Left: the system 'learning curve' (see Section 4.1.1 for full explanation) for the brain MR data. Right: interobserver distances (between observers 3 and 4) for the three sets of brain MR data, shown in full (above) and in closer view (below). For each of the three scan pairs the point distances are divided into manually matched points (left) and automatically matched points (right). The lightest grey colour indicates the boundaries of the central 50% of the data, mid-grey the central 75% and the darkest grey the central 90%. The horizontal line within the central 50% marks the median value. Outliers are included in the whisker length.

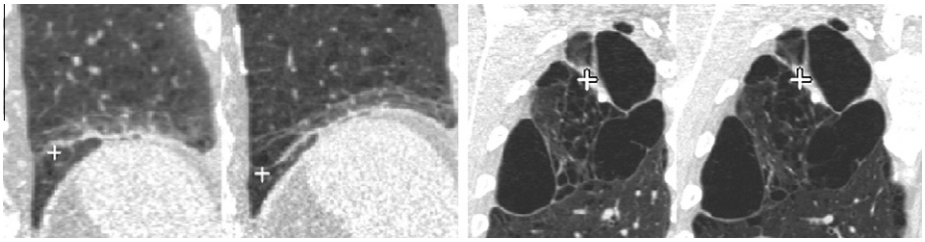

Fig. 20. Point pairs from scan pair 5 (left images) and scan pair 20 (right images). In each case the landmark location is shown on the left and the matching point from an observer on the right. In scan pair 5 neither observer nor the expert could find a good match for the landmark shown (note the point was not used in the evaluation due to interobserver differences >1 mm). In scan pair 20 a good match has been found for this point, although the scan is generally difficult due to severe emphysema. Marker sizes have been increased for visualisation in these images.

matching have not yet been studied and more significant system modifications may be required to achieve the same level of success with such data. In addition, it must be noted that there are registration tasks where it is extremely difficult for even an expert observer to make reliable point correspondences because of the nature of the image data or the anatomy being imaged. This is particularly true for images where very little structural detail can be seen. Our system for reference standard construction is clearly unsuited to such registration tasks.

The efficacy of the system in quantitative analysis of registration schemes is demonstrated by application to several sets of registration results, showing the distinctions between various methods. The affine-only (AF) and no-masks (NM) configurations are included largely as proof of concept since we clearly expect the results from these registrations to be inferior. The assessment based on the constructed reference standard (see Fig. 16) confirms that these registrations do give particularly poor results compared to the other methods. The basic registration (BS) was included as it had previously been experimentally determined to be relatively accurate and efficient at registering this type of data. Alternative methods varying the cost function (CC and MS) and the initial image resolution (FR) were added in order to determine their possible effects on the speed and accuracy of the registration procedure.

Fig. 16 illustrates minor distinctions between the registration results averaged over all scan pairs. Such subtle differences between registration algorithms may easily be overlooked by evaluation techniques based on segmentation overlap measures, small numbers of landmark locations or synthetically produced registration problems. It should be noted however that one limitation of this system of evaluation is that the reference standard may show a bias in favour of registration algorithms which themselves are based around a TPS scheme.

The selection of a suitable registration method, while not the focus of this work, is a topic which inevitably arises on studying the presented results. The optimal choice depends largely on the ultimate purpose of the registration and there are many complicating factors such as the degree of accuracy required and the restrictions on processing time. We therefore present only a brief and general discussion based on the data included in this work. Fig. 16 shows that on average the full-resolution registration (FR) appears to give the best results, although it is only a marginal improvement compared to the basic registration on down-sampled data (BS) or the registration using normalized cross-correlation (CC). The BS and CC methods are, however, approximately three to four times faster than the FR registration (see Table 3) and may therefore be more suitable in cases where computation time is of importance. The mean-squares (MS) registration is overall slightly less accurate but also slightly faster than the other non-rigid registrations using lung masks.

From the box plots depicted in Fig. 18 however, it is clear that a particular method may produce satisfactory results in one case and yet demonstrate serious inaccuracies in another, even when all datasets have the same general properties. A much greater range of differences between the algorithms is illustrated in this figure. The MS registration procedure gives generally inferior results to the FR method but for scan pairs 10 and 11, for example, the performance of both methods is very similar. Conversely, scan pair 8 is seen to be reasonably well registered by the BS, CC and FR methods, but is relatively poorly aligned by the MS technique. This information illustrates the fact that comparisons of registration algorithms based on results from different datasets are flawed and unreliable. In order to compare registration techniques in a meaningful way a large and diverse set of publicly available data is required.

Some scan pairs appear to have consistently poorer registration results than others across all the tested methods. Scan pairs 5 and 20, for example, both exhibit relatively large errors in all methods shown in Fig. 18. Upon further investigation, it was found that both of these scans exhibit pathology which is not seen in many subjects in this dataset since they are drawn from a screening trial. Subjects with pathology are much more likely to have tissue changes over time which are extremely difficult to handle with registration because of the appearance or disappearance of structures. Fig. 20 shows examples of points in each of these scan pairs.

Overall, for consistent and reliable registration results FR is clearly the best option among those tested with a median error of approximately 0.5 mm in the majority of scans. It should be noted that since the accuracy of the reference standard data is limited by the voxel size ($\approx$0.7 mm) it is not possible to evaluate a sub-voxel accuracy registration algorithm without detecting some degree of apparent error.

6. Conclusion

A semi-automatic scheme has been presented which enables the provision of extensive and accurate reference standard data for registration. The method has been demonstrated to work well on temporal chest CT data with both real and synthetic warping. It also performs well on brain MR data and has potential to achieve excellent results with some system modifications. It has been shown that the annotations of non-expert observers made with this system do not differ significantly from those of a radiology expert. An approach such as this, which is efficient and accurate is essential in order to comprehensively evaluate and detect subtle differences between the ever increasing number of registration algorithms under development.

Acknowledgements

The authors would like to thank Dr. Koen Vincken for making the MR data available.

References

Betke, M., Hong, H., Ko, J.P., 2003. Landmark detection in the chest and registration of lung surfaces with an application to nodule registration. Med. Image Anal. 7, 265–281.

Blaffert, T., Wiemker, R., 2004. Comparison of different follow-up lung registration methods with and without segmentation. In: Proceedings of the SPIE, vol. 5370, pp. 1701–1708.

Boldea, V., Sharp, G.C., Jiang, S.B., Sarrut, D., 2008. 4D-CT lung motion estimation with deformable registration: quantification of motion nonlinearity and hysteresis. Med. Phys. 35 (3), 1008–1018.

Bookstein, F.L., 1989. Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans. PAMI 11, 567–585.

Brown, L.G., 1992. A survey of image registration techniques. ACM Comput. Surveys 24, 325–376.

Castillo, R., Castillo, E., Guerra, R., Johnson, V.E., McPhail, T., Garg, A.K., Guerrero, T., 2009. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Phys. Med. Biol. 54 (7), 1849–1870.

Christensen, G.E., Geng, X., Kuhl, J.G., Bruss, J., Grabowski, T.J., Pirwani, I.A., Vannier, M.W., Allen, J.S., Damasio, H., 2006. Introduction to the non-rigid image registration evaluation project (NIREP). In: Third International Workshop on Biomedical Image Registration. Lecture Notes in Computer Science, vol. 4057. Springer-Verlag, Berlin, pp. 128–135.

Crum, W.R., Rueckert, D., Jenkinson, M., Kennedy, D., Smith, S.M., 2004. A framework for detailed objective comparison of non-rigid registration algorithms in neuroimaging. Med. Image Comput. Comput. Assist. Interv. 3216, 679–686.

Davis, M.H., Khotanzad, A., Flamig, D.P., Harms, S.E., 1997. A physics-based coordinate transformation for 3-D image matching. IEEE Trans. Med. Imag. 16 (3), 317–328.

Frantz, S., Rohr, K., Siegfried Stiehl, H., 2005. Development and validation of a multi-step approach to improved detection of 3D point landmarks in tomographic images. Image Vis. Comp. 23 (11), 956–971.

Glatard, T., Pennec, X., Montagnat, J., 2006. Performance evaluation of grid-enabled registration algorithms using bronze-standards. Med. Image Comput. Comput. Assist. Interv. 9 (Pt 2), 152–160.

Grachev, I.D., Berdichevsky, D., Rauch, S.L., Heckers, S., Kennedy, D.N., Caviness, V.S., Alpert, N.M., 1999. A method for assessing the accuracy of intersubject registration of the human brain using anatomic landmarks. Neuroimage 9 (2), 250–268.

Heath, E., Collins, D.L., Keall, P.J., Dong, L., Seuntjens, J., 2007. Quantification of accuracy of the automated nonlinear image matching and anatomical labeling (ANIMAL) nonlinear registration algorithm for 4D CT images of lung. Med. Phys. 34 (11), 4409–4421.

Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Le Goualher, G., Collins, D.L., Evans, A., Malandain, G., Ayache, N., Christensen, G.E., Johnson, H.J., 2003. Retrospective evaluation of intersubject brain registration. IEEE Trans. Med. Imag. 22 (9), 1120–1130.

Hill, D.L., Batchelor, P.G., Holden, M., Hawkes, D.J., 2001. Medical image registration. Phys. Med. Biol. 46, R1–R45.

Hu, S., Hoffman, E.A., Reinhardt, J.M., 2001. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. IEEE Trans. Med. Imag. 20, 490–498.

Jannin, P., Fitzpatrick, J.M., Hawkes, D.J., Pennec, X., Shahidi, R., Vannier, M.W., 2002. Validation of medical image processing in image-guided therapy. IEEE Trans. Med. Imag. 21, 1445–1449.

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

Kohlrausch, J., Rohr, K., Siegfried Stiehl, H., 2005. A new class of elastic body splines for nonrigid registration of medical images. J. Math. Imag. Vis. 23, 253–280.

Lester, H., Arridge, S.R., 1999. A survey of hierarchical non-linear medical image registration. Patt. Recognit. 32, 129–149.

Likar, B., Pernuš, F., 1999. Automatic extraction of corresponding points for the registration of medical images. Med. Phys. 26, 1678–1686.

Maintz, J.B.A., Viergever, M.A., 1998. A survey of medical image registration. Med. Image Anal. 2, 1–36.

Pevsner, A., Davis, B., Joshi, S., Hertanto, A., Mechalakos, J., Yorke, E., Rosenzweig, K., Nehmeh, S., Erdi, Y.E., Humm, J.L., Larson, S., Ling, C.C., Mageras, G.S., 2006. Evaluation of an automated deformable image matching method for quantifying lung motion in respiration-correlated CT images. Med. Phys. 33 (2), 369–376.

Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A., 2003. Mutual-information-based registration of medical images: a survey. IEEE Trans. Med. Imag. 22, 986–1004.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imag. 18 (8), 712–721.

Saff, E.B., Kuijlaars, A.B.J., 1997. Distributing many points on a sphere. Math. Intell. 19, 5–11.

Schnabel, J.A., Tanner, C., Castellano-Smith, A.D., Degenhard, A., Leach, M.O., Rodney Hose, D., Hill, D.L.G., Hawkes, D.J., 2003. Validation of nonrigid image registration using finite-element methods: application to breast MR images. IEEE Trans. Med. Imag. 22 (2), 238–247.

Simons, P.C., Algra, A., vande Laak, M.F., Grobbee, D.E., vander Graaf, Y., 1999. Second manifestations of arterial disease (smart) study: rationale and design. Eur. J. Epidemiol. 15 (9), 773–781.

Škerl, D., Likar, B., Pernuš, F., 2008. A protocol for evaluation of similarity measures for non-rigid registration. Med. Image Anal. 12 (1), 42–54.

Sluimer, I.C., Prokop, M., van Ginneken, B., 2005. Towards automated segmentation of the pathological lung in CT. IEEE Trans. Med. Imag. 24 (8), 1025–1038.

Thévenaz, P., Unser, M., 2000. Optimization of mutual information for multiresolution image registration. IEEE Trans. Image Process. 9, 2083–2099.

Tomazevic, D., Likar, B., Pernuš, F., 2004. "Gold standard" data for evaluation and comparison of 3D/2D registration methods. Comput. Aid. Surg. 9, 137–144.

Unser, M., 1999. Splines: a perfect fit for signal and image processing. IEEE Signal Process. Mag. 16, 22–38.

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

van de Kraats, E.B., Penney, G.P., Tomazevic, D., van Walsum, Th., Niessen, W.J., 2005. Standardized evaluation methodology for 2D–3D registration. IEEE Trans. Med. Imag. 24, 1177–1190.

Vandemeulebroucke, J., Sarrut, D., Clarysse, P., 2007. The POPI model, a point-validated pixel-based breathing thorax model. In: XVth International Conference on the Use of Computers in Radiation Therapy.

Vik, T., Kabus, S., von Berg, J., Ens, K., Dries, S., Klinder, T., Lorenz, C., 2008. Validation and comparison of registration methods for free breathing 4D lung CT. In: Proceedings of the SPIE.

Wang, H., Dong, L., O'Daniel, J., Mohan, R., Garden, A.S., Kian Ang, K., Kuban, D.A., Bonnen, M., Chang, J.Y., Cheung, R., 2005. Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy. Phys. Med. Biol. 50 (12), 2887–2905.

West, J., Fitzpatrick, J.M., Wang, M.Y., Dawant, B.M., Maurer Jr., C.R., Kessler, R.M., Maciunas, R.J., Barillot, C., Lemoine, D., Collignon, A., Maes, F., Suetens, P., Vandermeulen, D., vanden Elsen, P.A., Napel, S., Sumanaweera, T., Harkness, B., Hemler, P.F., Hill, D.L.G., Hawkes, D.J., Studholme, C., Maintz, J.B.A., Viergever, M.A., Malandain, G., Pennec, X., Noz, M.E., Maguire Jr., G.Q., Pollack, M., Pelizzari, C.A., Robb, R.A., Hanson, D., Woods, R.P., 1997. Comparison and evaluation of retrospective intermodality image registration techniques. J. Comput. Assist. Tomogr. 21, 554–566.

Wiemker, R., deHoop, B., Kabus, S., Gietema, H., Opfer, R., Dharaiya, E., 2008. Performance study of a globally elastic locally rigid matching algorithm for follow-up chest CT. In: Proceedings of the SPIE, vol. 6917.

Wörz, S., Rohr, K., 2006. Localization of anatomical point landmarks in 3D medical images by fitting 3D parametric intensity models. Med. Image Anal. 10 (1), 41–58.

Wörz, S., Rohr, K., 2006. New approximating gaussian elastic body splines for landmark-based registration of medical images. In: Bildverarbeitung Für Die Medizin.

Wu, Z., Rietzel, E., Boldea, V., Sarrut, D., Sharp, G.C., 2008. Evaluation of deformable registration of patient lung 4DCT with subanatomical region segmentations. Med. Phys. 35 (2), 775–781.

Xu, D.M., Gietema, H., deKoning, H., Vernhout, R., Nackaerts, K., Prokop, M., Weenink, C., Lammers, J., Groen, H., Oudkerk, M., van Klaveren, R., 2006. Nodule management protocol of the NELSON randomised lung cancer screening trial. Lung Cancer 54 (2), 177–184.

Zitová, B., Flusser, J., 2003. Image registration methods: a survey. Image Vis. Comput. 21, 977–1000.