Physics in Medicine & Biology

Institute of Physics and Engineering in Medicine

A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets

Phys. Med. Biol. **54** (2009) 1849–1870 doi:10.1088/0031-9155/54/7/001

A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets

Richard Castillo[1], **Edward Castillo**[2], **Rudy Guerra**[3], **Valen E Johnson**[4], **Travis McPhail**[5], **Amit K Garg**[6] **and Thomas Guerrero**[6,7]

[1] Department of Imaging Physics, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA
[2] Department of Mathematics, University of California, Irvine, CA, USA
[3] Department of Statistics, Rice University, Houston, TX, USA
[4] Department of Biostatistics & Applied Mathematics, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA
[5] Department of Computer Science, Rice University, Houston, TX, USA
[6] Department of Radiation Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

E-mail: tguerrero@mdanderson.org

Received 25 August 2008, in final form 18 January 2009
Published 5 March 2009
Online at stacks.iop.org/PMB/54/1849

Abstract

Expert landmark correspondences are widely reported for evaluating deformable image registration (DIR) spatial accuracy. In this report, we present a framework for objective evaluation of DIR spatial accuracy using large sets of expert-determined landmark point pairs. Large samples (>1100) of pulmonary landmark point pairs were manually generated for five cases. Estimates of inter- and intra-observer variation were determined from repeated registration. Comparative evaluation of DIR spatial accuracy was performed for two algorithms, a gradient-based optical flow algorithm and a landmark-based moving least-squares algorithm. The uncertainty of spatial error estimates was found to be inversely proportional to the square root of the number of landmark point pairs and directly proportional to the standard deviation of the spatial errors. Using the statistical properties of this data, we performed sample size calculations to estimate the average spatial accuracy of each algorithm with 95% confidence intervals within a 0.5 mm range. For the optical flow and moving least-squares algorithms, the required sample sizes were 1050 and 36, respectively. Comparative evaluation based on fewer than the required validation landmarks results in misrepresentation of the relative spatial accuracy. This study demonstrates that landmark pairs can be used to assess DIR spatial accuracy within a narrow uncertainty range.

[7] Author to whom any correspondence should be addressed.

0031-9155/09/071849+22$30.00 Printed in the UK

1. Introduction

The translation of image processing research into clinical application is generally confounded by the lack of established methodology for validating new algorithms (Gee 2000, Lehmann 2002). For objective and clinically relevant evaluation, considerable attention must be paid regarding the selection of an appropriate reference standard upon which to base algorithm performance and determine clinical utility. For validation of deformable image registration (DIR), a number of reference standards have been utilized, including synthetically deformed images (Lu *et al* 2004, Guerrero *et al* 2004, Wang *et al* 2005b), high-contrast phantoms (Wang *et al* 2005b) and expert-delineated control points (Rietzel and Chen 2006, Brock *et al* 2005, Kaus *et al* 2007, Boldea *et al* 2008, Brock *et al* 2008). While synthetic images and phantoms might provide useful qualitative evaluation of DIR performance characteristics, they lack sufficient realism to provide credible validation of registration spatial accuracy for use in the clinical setting (Fitzpatrick 2001). The best standard, therefore, is one derived from actual patient image data, for which ground truth is not known.

The relative abundance of high-contrast, anatomical landmarks such as vessel and bronchial bifurcations make thoracic 4D CT image data particularly well suited for the manual tracking of prominent image features across multiple image volumes. Tracking such features offers a means for estimating the true transformation and provides measures for statistical analysis of DIR spatial accuracy. Recent published landmark-based validation studies of thoracic DIR reflect this notion (Brock *et al* 2005, Kaus *et al* 2007, Boldea *et al* 2008, Coselmon *et al* 2004, Keall *et al* 2005, Pevsner *et al* 2006, Sarrut *et al* 2006, Al-Mayah *et al* 2008, Wu *et al* 2008, Wolthaus *et al* 2008, Li *et al* 2008). To date, however, there is not a common standard or framework for either generating or utilizing the reference samples used to characterize DIR performance. As a result, a large range of reference sample sizes, with equally varying spatial distributions, have been used to validate novel DIR algorithms. This inconsistency in evaluation standards complicates the interpretation of individual validation studies and makes objective comparison of reported DIR spatial accuracies difficult and potentially misleading.

For thorough and unbiased characterization of DIR spatial accuracy performance, it is necessary to ensure that the validation landmark sets adequately sample the volume of interest not only spatially but also in terms of the clinically relevant variables that could potentially affect DIR output. Such factors include physiological motion characteristics such as displacement magnitude and hysteresis, image quality and intensity characteristics such as local contrast and change in intensity between images. With this in mind, it is important to distinguish between quantitative assessment for characterization or acceptance testing, as opposed to quality assurance (QA) purposes. In the former, the goal is to construct a complete description of the DIR performance characteristics, and in the case of acceptance testing, utilizing as much of the available information as necessary to provide an informed assessment regarding the routine clinical feasibility and potential shortcomings of a given DIR algorithm. Thus, landmark samples should be selected of sufficient size to facilitate statistical analysis of DIR spatial accuracy performance. Though the selection of the necessarily large validation landmark sets is not feasible for routine QA purposes, it is also presumably not necessary, provided a thorough evaluation of the algorithm was performed prior to routine clinical implementation. For QA, the goal is rather to ensure for any given case that the DIR spatial accuracy meets accepted standards within the context of the specific clinical application. The insight acquired during the characterization process is therefore crucial, and directly applicable to the development of specific QA testing procedures, that, based on only a limited amount of information for any given case, will nonetheless ultimately be used to judge the quality of the output in order to prevent potentially harmful errors from reaching the patient.

The goal of this study is to demonstrate a proposed consistent and self-contained framework for the objective evaluation of thoracic DIR. This framework is based on the use of large samples of expert-determined landmark feature pairs between volumetric images as a reference for spatial accuracy measurements, for purposes of

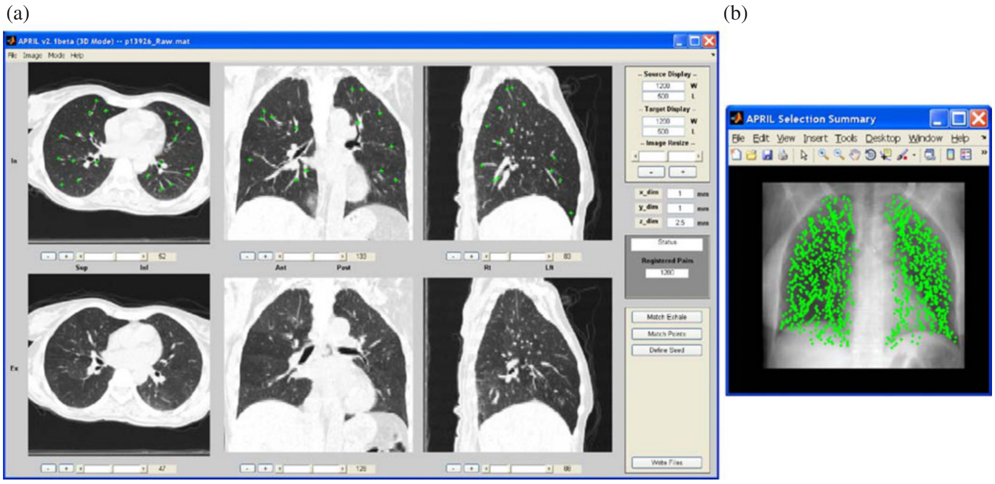
(a) optimization and characterization of DIR output during algorithm development,
(b) comparative evaluation of multiple DIR algorithms,
(c) formal acceptance testing of individual algorithms for specific clinical application and
(d) QA of DIR output in the routine clinical setting.

For landmark selection, we employ a novel Matlab-based (Mathworks, Sunnyvale, CA) software interface, developed to streamline the manual selection process and manage the corresponding samples of validation point sets. Using the interface, large samples ($>$1100) of corresponding pulmonary landmark features were manually generated from treatment planning 4D CT data to facilitate statistical evaluation of DIR spatial accuracy. In order to demonstrate the practical utility of the landmark sets for validation and comparative evaluation, we compare the spatial accuracy performance of two DIR algorithms, a gradient-based optical flow algorithm and a landmark interpolation algorithm based on moving least-squares, for registration of thoracic CT image pairs. Furthermore, we investigate the correlation of standard image intensity-based measures for assessing DIR performance with the spatial accuracy derived from the validation landmark sets. Finally, we utilize the statistical properties of the DIR output over the validation point sets to demonstrate the effect of landmark sample size on the uncertainty associated with calculated values for mean registration error.

The organization of the remainder of this document is as follows. Section 2 describes the process of generating the large samples of manually registered feature points for objective evaluation of DIR spatial accuracy. It is broken into five sections. Section 2.1 briefly describes the five clinically acquired patient images utilized throughout this study. Section 2.2 describes the experimental methods for landmark registration and section 2.3 describes the large landmark datasets generated from the five clinically acquired treatment planning 4D CT image volumes. Section 2.4 describes the statistical characterization of the landmark sets for uncertainties associated with observer variance, while section 2.5 addresses the issue of landmark localization uncertainty with regard to image resolution and voxel dimension. Section 3 demonstrates the practical utility and necessity of the large point sets both for characterization and comparative evaluation of DIR outputs. Sections 3.1 and 3.2 briefly describe the two DIR algorithms that are used in this study to generate example DIR datasets for the five patient cases. Section 3.3 focuses on the spatial accuracy characterization of both output sets derived from the validation landmarks. Additionally, in section 3.3 we investigate the correlation of standard image-intensity-based measures for assessing DIR performance with the spatial accuracy measurements derived from the validation landmark sets. Section 4 focuses on the statistical requirements on landmark sample size, with an example presented in the context of comparative evaluation of multiple DIR algorithms. Finally, section 5 summarizes the framework and provides a general discussion regarding its use.

2. Landmark selection and characterization

While it has been demonstrated in the literature that manually registered landmarks can be a useful tool for obtaining spatial accuracy measurements for DIR, a consistent framework for



-- Source Display --
1200 W
500 L
-- Target Display --
1200 W
500 L
-- Image Resize --
x_dim 1 mm
y_dim 1 mm
z_dim 2.5 mm
Status
Registered Pairs 1280

Figure 1. Manual registration interface. (a) Primary display window, showing source (top panel) and target (bottom panel) component phase volumes from a 4D CT set. Green crosshairs indicate manually registered source feature points. (b) Selection summary window displays a digitally reconstructed AP projection of the source image data. The display is updated as source points are selected to show a projection of the current set of sources for reference. The illustration depicts an image pair for which 1280 pulmonary landmarks have been manually delineated.

evaluation is still lacking. This section describes our methodology for the generation of the manually registered point sets. For demonstration purposes, five clinically acquired patient datasets were obtained.

2.1. Thoracic CT images

The treatment planning 4D CT images from five patients free of pulmonary disease who were treated for esophageal cancer were selected. Patient identifiers were removed in accordance with an institutional review board approved retrospective study protocol (RCR 03-0800). Each patient underwent treatment planning in which 4D CT images of the entire thorax and upper abdomen were acquired at 2.5 mm slice spacing with a General Electric Discovery ST PET/CT scanner (GE Medical Systems, Waukesha, WI). The extreme inhale and exhale phases of the 4D CT image sets were utilized in this study. Each image was cropped to include the entire rib cage and content sub-sampled to 256 × 256 voxels. Final in-plane voxel dimensions ranged from (0.97 × 0.97) to (1.16 × 1.16) mm^2. No sub-sampling was performed in the superior–inferior direction. For all cases, the final image slice thickness was 2.5 mm.

2.2. Landmark selection

A Matlab-based software interface named APRIL (*A*ssisted *P*oint *R*egistration of *I*nternal *L*andmarks) was developed in-house to facilitate manual selection of landmark feature pairs between volumetric images. Figure 1 shows the main interface display. Two volumes are simultaneously displayed in transverse, coronal and sagittal orientation in the primary display window (figure 1(a)) with the top panel fixed to display the designated source volume. Basic features of the software include separate window and level settings for each display, visualization of equivalent voxel locations in the orthogonal planes and interactive tools for segmentation of lung voxels from the image data.

Manual registration of the volumetric image pair begins with the selection of a unique feature point within the designated source volume via mouse click on any of the orthogonal source displays. Upon selection, the feature voxel is highlighted for reference in each of the source and target images. To assist in the manual registration process, the software provides an optional feature localization tool based on normalized cross-correlation of a size-adjustable local voxel neighborhood (Lewis 1995, Gonzalez and Woods 2008). Given the feature voxel v located at position (S_x, S_y, S_z) in the source image, we create an $(m \times m \times m)$ source neighborhood, N_S, centered on v. The isotropic neighborhood dimension is given by $m = 2\alpha + 1$, where the parameter α is chosen by the user from a list of available dimensions that range from 4 to 32 voxels. A user-defined intensity threshold is applied to N_S to generate a binary mask of the original source neighborhood, designated N_S^*. A local neighborhood of the same dimension is similarly defined, centered on v in the target volume, and the user threshold applied to generate N_T^*. The normalized cross-correlation coefficient, δ, at location (x, y, z) in the target image, is then given by

$$\delta(x,y,z) = \left\{ \sum_{r=-\alpha}^{\alpha} \sum_{s=-\alpha}^{\alpha} \sum_{t=-\alpha}^{\alpha} [N_S^*(S_x + r, S_y + s, S_z + t) - \bar{N}_S^*][N_T^*(x + r, y + s, z + t) - \bar{N}_T^*] \right\} \Bigg/ \left\{ \sum_{r=-\alpha}^{\alpha} \sum_{s=-\alpha}^{\alpha} \sum_{t=-\alpha}^{\alpha} [N_S^*(S_x + r, S_y + s, S_z + t) - \bar{N}_S^*]^2 \times \sum_{r=-\alpha}^{\alpha} \sum_{s=-\alpha}^{\alpha} \sum_{t=-\alpha}^{\alpha} [N_T^*(x + r, y + s, z + t) - \bar{N}_T^*]^2 \right\}^{0.5}, \tag{1}$$

where $\bar{N}_S^*$ and $\bar{N}_T^*$ are the average intensity values within the respective binary source and target neighborhood masks. To facilitate rapid feature localization, δ is only calculated over the region shared by N_S and the target volume. By using only the binary source and target neighborhoods, cross-correlation is performed only over the local structural content, where the level of included structural detail is controlled via the user-defined intensity threshold. The target voxel representing the maximum of the 3D correlation function is highlighted in the target displays and represents an estimate of the feature correspondence. In practice, multiple correlations varying both neighborhood dimensions and/or intensity threshold may be performed. However, the user ultimately must manually designate the feature correspondence via mouse click on the target image. Following confirmation of the target selection, the process is repeated until the desired sample size and uniformity of distribution have been achieved.

A selection summary window (figure 1(b)) is actively updated to display a plot of the current registered source locations, projected onto a digitally reconstructed anterior–posterior (AP) projection of the source image content. The orthogonal source displays are also updated to illustrate all accepted source voxels on the current slice (if any). This is useful for assessing adequate feature distribution on a slice-by-slice basis and also prevents redundant selection of source landmarks. Following manual registration, summary text files are exported to streamline analysis procedures. The summary includes Cartesian and spherical coordinate lists of the source and target feature locations, corresponding voxel intensities and displacement magnitudes (separately, in units of millimeters and voxels).

2.3. Landmark datasets

Pulmonary landmark feature pairs, typically vessel bifurcations, were manually delineated on the five test image pairs by an expert in thoracic imaging. Source feature points were selected

Table 1. Reference landmark characteristics. The number of expert-determined landmark feature points is shown for each case in terms of right, left and total lung points. The number of landmarks that were not displaced ($D = 0$) between the maximum inhale and exhale 4D CT phases is also shown for each case. Average and maximum landmark displacements are seen to vary substantially across the five datasets.

	Case 1	Case 2	Case 3	Case 4	Case 5
	Validation landmark points				
# Right lung	672	767	637	803	661
# Left lung	608	720	529	758	607
Total	1280	1487	1166	1561	1268
$D = 0$	85	189	11	10	35
Average displacement (SD)[a] (mm)	4.01 (2.91)	4.65 (4.09)	9.42 (4.81)	6.73 (4.21)	7.10 (5.14)
Maximum displacement (mm)	12.65	17.8	21	18.46	24.78

[a] Standard deviation.

systematically, beginning at the apex of the lung, with an initial goal of >10 feature points for each lung per axial image slice. This approach ensured the collection of >1100 validation point pairs for each case. Following feature selection for a given case, all landmark pairs were visually reviewed by the primary reader a second time and the location adjusted on the exhale image if necessary. The verification step was required before the initial registration process performed by the primary reader was considered complete.

The number of registered feature pairs per case ranged from 1166 to 1561. A total of 6762 landmarks were manually registered over the set of five image pairs. On average, approximately 12 h, distributed over multiple sessions, were required to register a single case. Characteristics of the landmark pairs are summarized in table 1. Average displacement (and standard deviation) of registered features per case ranged from 4.01 (2.91) to 9.42 (4.81) mm, while maximum landmark displacements ranged from 12.65 to 24.78 mm. Average magnitude displacements in component right–left (RL), anterior–posterior (AP) and superior–inferior (SI) directions ranged from 0.58 (0.62) to 1.17 (1.05) mm, 0.67 (0.79) to 1.74 (1.67) mm and 3.68 (3.03) to 8.98 (5.04) mm, respectively. Figure 2 shows vector plots of the landmark displacement fields for the five cases in anterior (top row) and lateral (bottom row) projection. The sampled feature points are sufficiently distributed to capture the substantially heterogeneous spatial distribution of tissue motion within each of the lung volumes.

The stated goal of >10 feature pairs per lung per axial image slice served as a guideline to ensure uniform spatial distribution of the validation landmarks. In practice, the number of landmarks required to adequately sample a given image slice may be more or less, depending on the volume of lung contained within that slice. To demonstrate this effect, the inhale lung voxels for each case were segmented based on simple histogram segmentation and three-dimensional connectivity. The lung volumes were then partitioned into blocks, each approximately 1/8 of the total number of axial slices containing lung. Figure 3(a) shows the number of landmarks contained within each of the sectioned superior–inferior blocks for all cases, while figure 3(b) shows the corresponding distribution of total lung volume over the same superior–inferior extent. Though the volume measurements are only approximations based on image segmentation, figure 3 suggests that the quantity of selected feature points as a function of location in the superior–inferior direction is primarily attributable to the superior–inferior distribution of lung volume. Note that figure 3(b) does not provide any indication as to the relative lung volumes between cases.

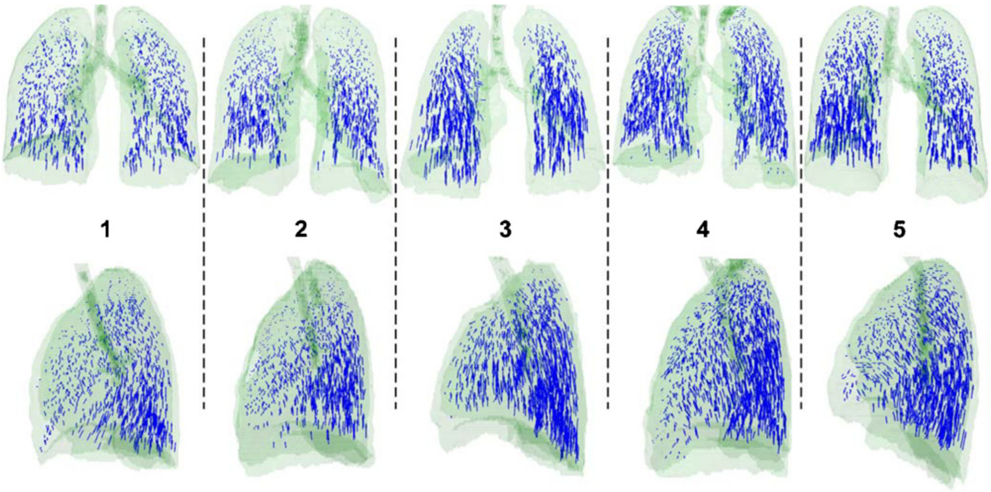


Figure 2. Validation landmark sets. Manually determined displacement vectors are shown in anterior (top row) and lateral (bottom row) projection for the five CT image pairs (case numbers are indicated). The base of each vector represents the location of a landmark feature in the maximum inhale phase of a 4D CT, while the head represents the corresponding feature location in the respective maximum exhale phase. 1280, 1487, 1166, 1561 and 1268 individual landmarks were manually selected for cases 1 through 5, respectively.

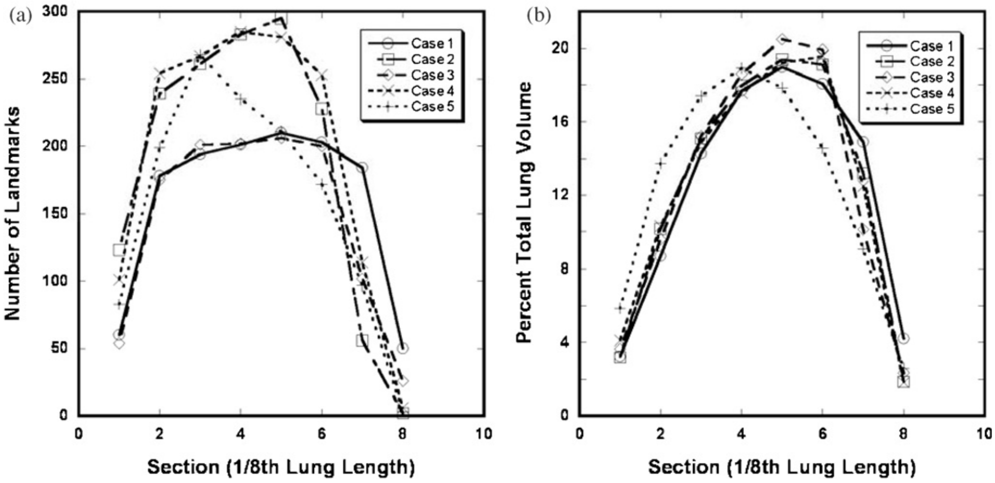


Figure 3. Superior–inferior distribution of landmark points and lung volume. Each lung volume was partitioned into blocks approximately 1/8 of the total lung length (i.e. total axial slices). For cases in which the number of lung slices was not divisible by 8, the remaining slices were included in the eighth section. (a) The cumulative number of landmarks is shown for each case as a function of the sectioned superior–inferior extent of the lung. (b) The percentage of total lung volume within each section is shown for all cases.

2.4. Landmark selection variability

In order to provide estimates of reproducibility of target point selections, random samples of 200 source feature points were generated for each case from the primary landmark sets. The source lists were then imported into the APRIL interface and re-registered by two secondary

Table 2. Landmark reproducibility summary. Inter- and intra-observer reproducibility of individual target point selections were estimated from repeated registration of uniform samples of 200 source points for each case. Mean errors were also determined for the set of repeated registration measurements. Two-sample *t*-tests were performed comparing mean inter- and intra-observer repeated registration errors (corresponding *p*-values shown).

	Case 1	Case 2	Case 3	Case 4	Case 5	All points
			Repeated registration errors (mm)			
Primary reader (SE)[a]	0.88 (0.07)	0.61 (0.07)	1.11 (0.07)	0.73 (0.07)	0.84 (0.08)	0.83 (0.03)
Secondary readers (SE)[a]	0.83 (0.07)	0.74 (0.05)	1.14 (0.07)	0.79 (0.05)	0.95 (0.06)	0.89 (0.03)
p	0.6473	0.1562	0.7617	0.4758	0.2691	0.2079

[a] Standard error.

readers to estimate inter-observer reproducibility. The primary reader also re-registered the sampled sources to estimate intra-observer reproducibility. Each of the repeated registrations was done independently and without prior knowledge of the primary target point selections.

The repeated registration error was quantified as the three-dimensional Euclidean distance between the original target point in the primary dataset and the corresponding point selected in the repeated registration. Repeated registration errors by the secondary readers were combined to estimate inter-observer reproducibility, while the intra-observer estimates were determined from repeated registrations by the primary reader. Mean errors and corresponding standard errors were calculated for each case, as well as over the combined set of error measurements. The observed error distributions across all cases for both the primary and secondary readers were skewed with respective skewness parameter values of 1.3 and 3.2. Two-sample *t*-tests were performed comparing mean inter- and intra-observer repeated registration errors. Although the error distributions were skewed, the Central Limit Theorem (Casella and Berger 2001) justifies a *t*-test for comparison of means. A summary of the error measurements is provided in table 2.

In four out of the five cases, mean intra-observer registration error was lower than mean inter-observer error, though the differences did not reach statistical significance for any of the cases ($p \geqslant 0.1562$). Mean repeated registration errors (SE) by the primary reader ranged from 0.61 (0.07) to 1.11 (0.07) mm for the five cases, while mean inter-observer errors ranged from 0.74 (0.05) to 1.14 (0.07). Over the combined set of 3000 repeated registration measurements, mean error for all observers was 0.87 (0.02) mm, with an inter-quartile range of 1.16 mm.

2.5. *Spatial localization uncertainty*

Finite sampling and image resolution due to acquisition and reconstruction inherently impose fundamental uncertainties associated with spatial localization of anatomical landmarks in medical images. For a fixed image resolution and voxel size, an observer cannot meaningfully localize a prominent feature point with sub-voxel accuracy. Thus, the manually determined landmark correspondences are described by integer coordinate pairs. In general, when quantifying landmark-based registration errors one must take into account that there is an inherent spatial uncertainty associated with the voxel localization of each landmark feature that is a function of the voxel dimension in each direction. For the CT images utilized in this study, the maximum RL and AP voxel dimensions were 1.16 mm. The SI voxel dimension for each case was 2.5 mm. However, because of the large number of landmarks included in our analysis, we were able to estimate the average error associated with the landmark identification by both the readers and the registration algorithms with sub-voxel accuracy. To understand why this is possible, recall that a Bernoulli proportion can be estimated with arbitrary precision (for a sufficiently large sample size), even though the outcome of each Bernoulli trial takes only one of two discrete values (0 and 1). Furthermore, we require as a specific criterion for the manual selection of point pairs that the image features are identifiable in both source and target images. In this sense, a secondary effect of image resolution on landmark selection is on the quantity of feature points satisfying this criterion for a given image pair. Thus, for relatively poor resolution images fewer usable landmarks can be identified.

3. Landmark-based evaluation of DIR

The goal of DIR is to find a point-to-point correspondence between two given images. This desired correspondence should relate the location of each underlying tissue element represented in each voxel in the first image to that in the second image. In order to demonstrate the utility of the large landmark sets as a means for assessing DIR performance, two DIR algorithms were implemented, providing example DIR output for the five patient cases described above. The two methods are briefly described.

3.1. Optical flow DIR

Optical flow methods (OFM) (Horn and Schunck 1981) comprise a large class of image registration techniques where the voxel correspondence is determined by computing a velocity field describing the apparent motion depicted in the two images. For a single pair of images, the velocity field is equal to the displacement field with the time step assumed to be unity. Several reviews of these methods exist, as do studies that focus on the performance of different optical flow implementations and techniques (see Beauchemin and Barron (1995), for example). In a previous work, we employed optical flow to track tumor motion and calculate ventilation from 4D CT (Guerrero *et al* 2004, 2006). Our optical flow implementation is based on an iterative procedure (Horn and Schunck 1981) used to solve for the unknown velocity at each voxel:

$$v_{n+1} = \bar{v}_n + \nabla I \left(\frac{\nabla I \cdot \bar{v}_n + \frac{\partial I}{\partial t}}{\alpha^2 + \|\nabla I\|^2} \right), \qquad (2)$$

where n and $n + 1$ are iteration counts and $\bar{v}_n$ is the average velocity taken over the nearest neighboring voxels. This method is equivalent to the well-known Gauss–Seidel method (Press *et al* 2002) where the latest available velocity values are used in calculating the average. All necessary temporal and spatial image derivatives are approximated with finite differences applied to the two given images. In this study, eight iterations of equation (2) with $\alpha = 25$ were performed for all OFM image registrations. To ensure variation in DIR spatial accuracy performance, no attempts were made to optimize individual case registrations.

3.2. Landmark-based DIR

Landmark-based algorithms represent an alternative class of image registration techniques in which sets of registered control point pairs are used to calculate an interpolating function that estimates the displacement of all voxels within the volume of interest. In this study, control point features in the inhale images are selected automatically by iteratively storing the locations of maximum intensity voxels in an edge-enhanced version of the original image data. Once selected, a spherical region surrounding the given control point is set to zero in the edge-enhanced image to ensure uniform spacing of the feature points. Estimates of the corresponding exhale landmarks are determined automatically according to a weighted cross correlation of the local source neighborhood with a larger search region in the target image. The weighting factor is included to scale the correlation function inversely as a function of distance from the source feature point.

The automated control point correspondences were visually assessed and modified as necessary prior to DIR. The interpolation step is performed using a moving least-squares (MLS) algorithm applied to the control points. Given the set of control point pairs, an affine function $A_v(x)$ is determined for each voxel v by minimizing the expression

$$\min \sum_i w_i \|A_v(p_i) - q_i\|^2, \tag{3}$$

where p_i and q_i are the ith source and target control point pair, respectively. The w_i are of the form

$$w_i = \frac{1}{\|p_i - v\|^2 + \varepsilon}, \tag{4}$$

where $\varepsilon > 0$. Thus, the interpolated value at v is given by $A_v(v)$. This implementation of MLS for landmark-based DIR has been previously described in detail (Schaefer *et al* 2006). In this study, a maximum of 368 interpolation landmarks were used for a single case. Over the set of five cases, a total of seven landmark features were common to the interpolation and validation landmark sets. For purposes of this study, the seven landmarks were included in the registration error analysis.

3.3. Evaluation of DIR

Fundamentally, registration error is defined as the difference between a calculated output and the designated reference standard. In this case, large sets of manually delineated feature pairs serve as the primary validation data. For this comparison to be strictly valid, the evaluation of manual and calculated landmark registration should be equivalent. That is, since an observer selects integer voxel locations in an image pair as corresponding point sets, the comparison with calculated positions should also be performed on the same integer grid. This is achieved simply by rounding the final displaced position of each coordinate of interest to the nearest integer. As described in section 2.5, we were able to estimate the average error associated with the landmark identification by the registration algorithms with sub-voxel accuracy, due to the large measurement sample sizes. Numerically, the mean errors determined from the rounded and floating point DIR positions will likely be similar. This is due to the fact that on average approximately equal quantities of test voxels are rounded toward their respective reference target position as are rounded away. However, to ensure equivalence of the reference standard and the calculated outputs, integer positions should be utilized.

Point registration error was quantified as the three-dimensional Euclidean distance between target voxels in the primary dataset, and those determined by applying the calculated DIR transformation to the corresponding source feature location. Mean registration error and corresponding standard error were determined for both DIR algorithms over the set of validation landmarks, providing a global measure of spatial accuracy performance for each case. Mean errors were also determined over the combined set of expert-determined feature points for all cases. Additionally, errors were assessed separately for individual RL, AP and SI component directions. Two-sample *t*-tests were performed in order to assess the statistical significance of differences in the mean registration error between algorithms. The skewness parameter (Casella and Berger 2001) values for OFM and MLS were 1.9 and 2.3, respectively. As above, the *t*-test was justified for comparison of the mean errors by the Central Limit Theorem (Casella and Berger 2001). For both methods, since the observed

Table 3. DIR spatial accuracy comparison. Three-dimensional and component mean registration errors derived from the complete validation point set are shown for both DIR algorithms. Two-sample *t*-tests were performed to assess the difference in mean errors, with corresponding *p*-values shown. Spearman's rank correlation coefficient was also determined to investigate the correlation of registration error with magnitude landmark displacement, local source feature contrast and change in landmark intensity between images. $p < 0.0001$ for all Spearman's coefficients except where indicated. All registration errors shown are in units of millimeters.

	OFM	MLS	*p*-value
	Mean error (SE)[a]		
3D	6.90 (0.10)	2.05 (0.02)	<0.0001
Right–left	2.21 (0.04)	0.77 (0.01)	<0.0001
Anterior–posterior	4.94 (0.09)	0.90 (0.01)	<0.0001
Superior–inferior	2.78 (0.05)	1.05 (0.02)	<0.0001
	Spearman's rank correlation		
Magnitude displacement	0.562	0.203	
Local source contrast	−0.100	−0.12	
Intensity change	0.107	0.014*	

[a] Standard error.

* $p = 0.2578$.

error distributions were skewed, the non-parametric Spearman rank correlation coefficient was calculated to quantify the statistical correlations between registration error and each of the following: displacement magnitude, change in intensity between image pairs and local contrast within a (5 × 5 × 5) voxel neighborhood surrounding each source feature point.

Table 3 summarizes the spatial accuracy performance of the two DIR algorithms for registration of the five thoracic CT image pairs. Over the complete validation landmark set, the mean registration errors (SE) for respective OFM and MLS DIRs were 6.90 (0.10) and 2.05 (0.02) mm. The inter-quartile ranges for the OFM and MLS DIR were 10.03 and 1.63 mm, respectively. For the OFM DIR, the mean RL, AP and SI component errors were each greater than 2 mm, with the largest registration errors occurring in the AP direction. In contrast, all MLS mean component errors were less than 2 mm, with the largest occurring in the SI direction. Spearman's rank correlation coefficients were calculated to assess the correlation of registration error with landmark displacement magnitude, change in intensity and local source landmark contrast. For both algorithms, the largest correlation was observed for displacement magnitude, with corresponding OFM and MLS Spearman's coefficients of 0.562 and 0.203, respectively.

The summary data presented in table 3 are useful for providing a global assessment of the spatial accuracy characteristics of the two DIR algorithms. However, a more detailed investigation of the nature of the errors can easily be performed through graphical interpretation of the same results by binning the error measurements appropriately. For example, the correspondence between registration error and landmark displacement is depicted graphically for both algorithms in figure 4. For each case, registration errors were binned according to the magnitude displacement of the complete set of 6762 validation landmarks in 2 mm increments. The figure shows relatively consistent behavior of the MLS registration errors over the range of displacement magnitudes. A positive trend can be seen for the OFM errors, consistent with the calculated Spearman correlation coefficients. This trend is not a strictly monotonic function of displacement magnitude, however, due to the fact that the OFM errors are more strongly a function of position rather than displacement (figure 6).

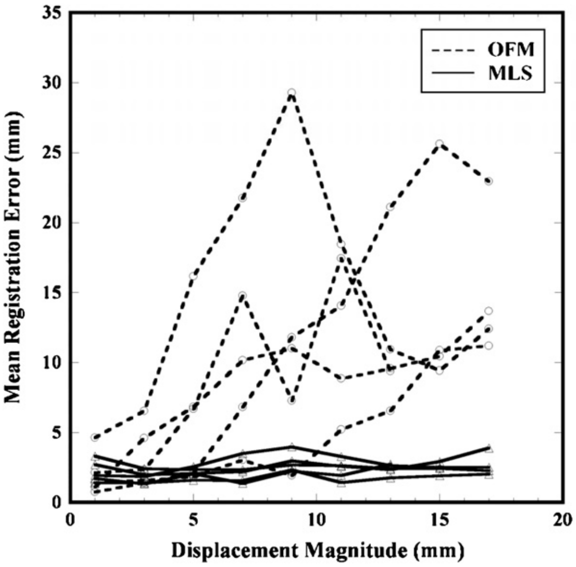


Figure 4. Registration error versus displacement magnitude. For each DIR algorithm, registration errors were binned corresponding to magnitude displacement of the validation landmarks in 2 mm increments. Mean registration errors were then determined for each bin and plotted versus displacement, separately for all five 4D CT image pairs.

Figure 5 shows a bar graph depiction of the mean registration errors for each individual case. The corresponding intra-observer repeated registration errors, as well as the mean landmark displacement magnitudes, are similarly shown for reference. The mean case errors ranged from 1.47 (0.03) to 2.55 (0.05) mm for MLS DIR and 3.73 (0.14) to 13.96 (0.38) mm for the OFM DIR. Those cases for which the DIR registration error is greater than the corresponding landmark displacement magnitude indicate instances in which deformable registration of the image pairs resulted in increased misalignment of the landmarks. It is important to note that there is no indication of this in the summary error statistics presented in table 3. Numerical spatial accuracy measurements can only be properly interpreted with reference to the validation data from which the measurements were acquired. Furthermore, reference should also be made to the inter- and intra-observer variance obtained during characterization of the landmark datasets. An algorithm that achieves a statistically indistinguishable result when compared to the expert landmarks effectively reaches the maximum resolution of the dataset.

Volume renderings of the lung surfaces were also generated and overlaid with a vector representation of each of the individual error measurements to visually assess the spatial distribution of registration errors within the anatomic context. Figure 6(a) shows the vector representation of the expert-determined validation landmark point set for an example case, projected onto a surface rendering of the corresponding inhale CT lung voxels. The residual error vectors for both DIR algorithms are shown in figures 6(b) and (c). In these figures, the error vectors point from the manually delineated feature location in the target image to that determined from the respective DIR transformation. The OFM plot shows a relatively large AP component error, consistent with the global assessment presented in table 3. Graphically, no systematic tendencies are apparent for the MLS DIR, suggesting little correlation of registration error with spatial location. More detailed graphical or quantitative error analyses within

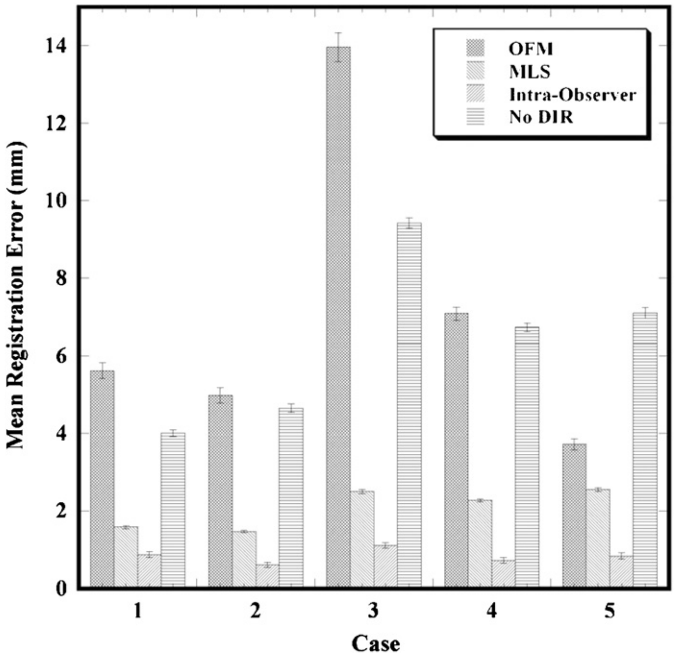


Figure 5. Case registration errors. Mean OFM and MLS DIR registration errors (± standard error) are shown for each case. Corresponding intra-observer repeated registration errors, as well as landmark displacement ('No DIR'), are also shown for reference. Those cases for which DIR registration error is greater than the corresponding landmark displacement magnitude indicate instances in which deformable registration of the image pairs resulted in increased misalignment of the validation landmarks.

specific regions of the lung volume, for example on an individual lung lobe basis, can be achieved simply by applying binary masks of the desired ROIs to the raw error measurement data.

Finally, a standard image intensity-based measure of DIR performance (Dawood *et al* 2008, Wang *et al* 2005a, 2005b, Lu *et al* 2004, 2006) was also determined for comparison with the spatial accuracy measurements derived from the validation point sets. For each case, estimated inhale image volumes were generated by applying the calculated DIR transforms to the inhale voxel grid and performing tri-linear interpolation of the mapped exhale neighborhood intensities to determine the estimated intensity of each voxel. A coronal slice from an example inhale image is shown in figure 7(a), next to the corresponding slice from the estimated inhale image derived from the optical flow DIR (figure 7(b)). The difference image is also shown in figure 7(c). Visually, the images appear similar. However, visual inspection alone provides no indication of the underlying DIR spatial accuracy. The mean registration error over the set of validation landmarks for the case depicted was 4.98 (SD: 7.66, Max: 41.76) mm.

Correlation coefficients were then calculated to assess the quantitative gray-scale similarity of the inhale and estimated inhale volumes. To avoid the influence of background voxel intensities and to determine correlation coefficients on an individual lung basis, the calculation was masked by separate right and left lung regions of interest (ROIs) determined from the original inhale image. Using the same lung ROIs, correlation coefficients were

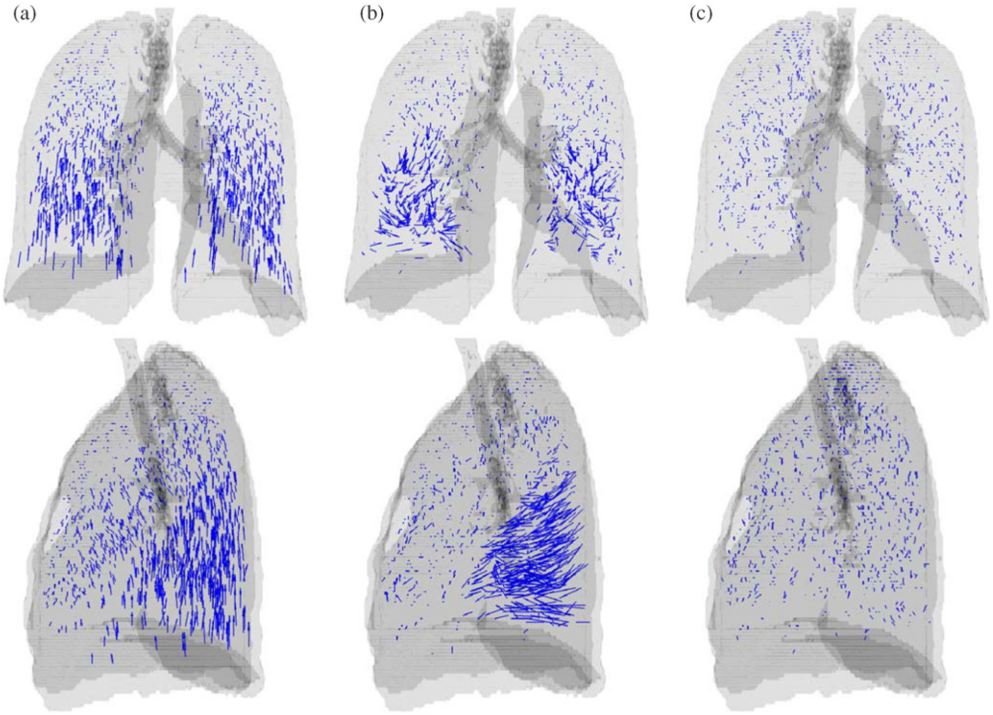


Figure 6. Residual error renderings. (a) 1487 pulmonary landmark features were manually registered between the maximum inhale and exhale component phase volumes from a 4D CT (case 2). The expert-determined displacement vectors are shown projected onto a surface rendering of the inhale lung volume. Residual error vectors are also shown for (b) OFM and (c) MLS. Each error vector points from the manually delineated feature location in the target image to that determined from the respective DIR transformation.

similarly calculated between the inhale and unregistered exhale volumes to determine the effect of DIR on image similarity.

In figure 7(d), the calculated changes in correlation coefficient were plotted versus the difference in mean registration error of the validation landmarks before and after DIR to graphically assess correspondence of the two performance metrics. In this example, a positive change in correlation coefficient indicates an increase in image similarity within the lung ROI following DIR, while a positive change in spatial error indicates an *increase* in misalignment of the validation landmarks. The two DIR algorithms exhibit different behavior; a greater increase in correlation was found with the OFM algorithm and a consistent reduction in spatial error was found with the MLS algorithm. For a majority of the test cases, the OFM algorithm resulted in an increase in the spatial error. The lack of correlation between image similarity and DIR spatial error is a new finding. For both algorithms, DIR consistently resulted in improved image similarity within the lung ROIs. However, the increase in correlation provided no indication of the underlying spatial error, which was made worse in some cases. For the data presented in this study, the correlation coefficient fails to provide even a reliable measure for relative performance between algorithms. For objective evaluation of the spatial accuracy of the calculated displacement of individual volume elements, intensity-based metrics afford little useful insight, as no information is provided regarding the origin of the aligned voxel

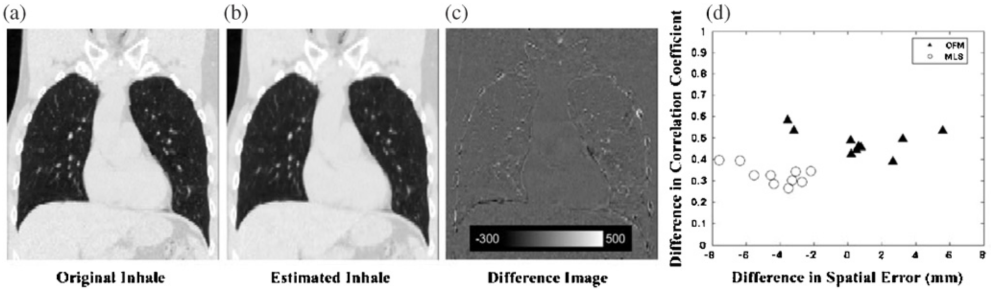


Figure 7. DIR performance metrics. Coronal CT slices are shown for an example case from (a) the original CT data next to (b) the corresponding slice from the estimated inhale image derived from the optical flow DIR. (c) The difference image is also shown. Visual and quantitative assessment of image similarity following DIR can result in potentially misleading evaluation of DIR spatial accuracy performance. For the volumetric image pair depicted, the mean registration error was 4.98 (SD: 7.66, Max: 41.76) mm. (d) Difference in correlation coefficient is shown versus corresponding difference in landmark registration error before and after DIR. Lung voxel ROIs were determined from the set of inhale images, separately for individual right and left lungs in order to increase measurement sample size. Positive change in correlation coefficient indicates increased image similarity following DIR, while negative change in spatial error indicates improved alignment of validation landmarks. Note that increased correlation coefficient does not necessarily imply improved spatial accuracy.

intensities (regardless of their equality). Hence, correlation and gray-scale similarity measures (Dawood *et al* 2008, Wang *et al* 2005a, 2005b, Lu *et al* 2004, 2006) and/or visual checks (Sharpe and Brock 2008) are inadequate for evaluation of DIR results.

4. Landmark sample size analysis

Objective evaluation of DIR based on large samples of landmark point sets can be a highly effective and informative strategy for characterization and comparative evaluation of algorithm performance. However, the large landmark datasets represent more than a useful tool when detailed assessment of DIR is desired. Rather, they should be considered a statistical necessity when the spatial accuracy characteristics of a given algorithm are not established *a priori*. In this section, we utilize the statistical properties of the two sets of DIR outputs over the validation point sets to demonstrate the effect of landmark sample size on the uncertainty associated with spatial error estimation.

For both algorithms, cumulative distribution functions (CDFs) were generated from the corresponding set of error measurements for each case. To simulate the spatial error information derived from validation point sets of different size, uniform samples of the individual CDFs were obtained for sample sizes ranging from 10 to 5000. For each sample size, 100 000 independent sample sets were obtained. At each sample size increment, an independent calculation of the mean spatial error was performed for each of the 100 000 error samples. The distribution of sample means was then used to determine the expected mean spatial error ±95% confidence intervals (CIs). Figure 8(a) shows the distribution of the 100 000 experimentally determined mean registration errors for the fixed validation sample size of 200 corresponding to the OFM output for an example case. The corresponding 95% CIs are also indicated. Figure 8(b) shows the corresponding measurements for sample sizes ranging from 10 to 5000. The set of CIs obtained through simulation was then compared with predicted intervals derived from basic statistics considerations. A thorough description of the

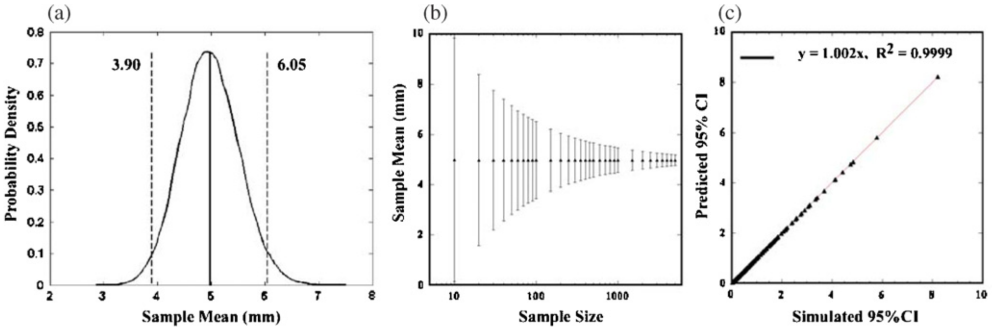


Figure 8. Mean registration error uncertainty. (a) The experimentally determined distribution of 100 000 mean registration errors is shown for an example case (case 5), for a fixed validation sample size of 200. Vertical bars indicate sample mean ± 95% CIs. (b) The distribution illustrated in (a) was similarly determined for validation sample sizes ranging from 10 to 5000. Corresponding sample means ± 95% CIs are shown in a semi-log plot, demonstrating the effect of sample size on the statistical uncertainty associated with mean registration error. (c) For all cases and both DIR algorithms, simulated versus predicted values for 95% CIs were plotted over the range of experimental sample sizes to assess linear correlation, with corresponding $R^2 = 0.99$.

statistical framework for uncertainty estimation can be found in most introductory textbooks in statistics or data analysis (for example, Bevington and Robinson (2003)). For each case and for both DIR algorithms, respective values for the standard deviation of error measurements were utilized to evaluate predicted values for the 95% CIs on the mean registration error for the experimental range of sample sizes described above. The combined set of predicted versus measured values for 95% CIs is plotted in figure 8(c). The square of the Pearson correlation coefficient, R, was calculated to assess linear correlation of the simulated and predicted CIs, with $R^2 = 0.99$, corresponding to the fitted linear regression equation given by $y = 1.002x$.

In practice, the statistical uncertainties associated with the mean registration error depicted in figure 8 can lead to potentially misleading assessment of DIR spatial accuracy characteristics. Figure 9 illustrates this point in the context of comparative evaluation of DIR spatial accuracy between two algorithms. For an example case, mean ± predicted 95% CIs are shown as a function of sample size for both OFM and MLS DIR. For this example, a minimum of approximately 150 uniformly distributed validation landmarks are required to obtain non-overlapping 95% CIs. Comparison of mean registration errors based on fewer than the required landmarks increases the probability that the comparative evaluation is a misrepresentation of the relative spatial accuracy characteristics of the two algorithms.

In general, the sample size required to obtain non-overlapping CIs will vary across test cases. Thus, for a given algorithm, it would be beneficial to formulate an estimate of the sample size required to obtain 95% CIs of a specified length (e.g. 1 mm) on the mean registration error. To do so requires incorporating all available error information pertaining to a given algorithm in order to calculate a pooled standard deviation of error measurements obtained from all available cases. The pooled standard deviation, $\bar{s}^{p}_{\mathrm{DIR}}$, is given by

$$\bar{s}^{p}_{\mathrm{DIR}} = \left[\left[\sum_{i=1}^{C}(N_i - 1)\right]^{-1}\left[\sum_{i=1}^{C}(N_i - 1)s^2_{i,\mathrm{DIR}}\right]\right]^{0.5}, \qquad (5)$$

where C is the total number of available cases, N_i is the validation sample size for the ith dataset and $s_{i,\mathrm{DIR}}$ is the corresponding standard deviation of the error measurements associated

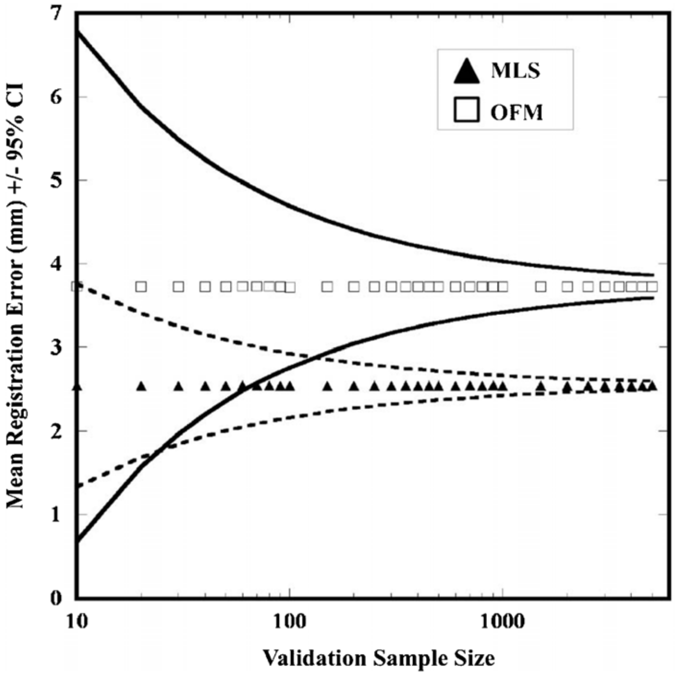


Figure 9. Comparative evaluation uncertainty. A semi-log plot of predicted uncertainty in mean registration error versus sample size is shown for an example case. Predicted values for 95% CIs were determined for validation sample sizes ranging from 10 to 5000. For both DIR algorithms, the 95% CIs on the mean registration error are shown. For this case, a sample size of approximately 150 validation landmarks is required to obtain non-overlapping 95% CIs. Comparative evaluation based on fewer than the required landmarks increases the probability that the comparison is a misrepresentation of the relative DIR spatial accuracy performance.

with a DIR algorithm. If we require a 95% CI within the specified range ($\pm d$ mm) of the mean error, then the necessary validation sample size is approximated by

$$N_{\pm d}^{\mathrm{DIR}} = \left(\frac{2\bar{s}_{\mathrm{DIR}}^{p}}{d}\right)^2. \tag{6}$$

This is an important point to consider in the interpretation of DIR spatial accuracies reported in the literature. The resulting sample size is not necessarily a large number (for example, for $\bar{s}_{\mathrm{DIR}}^{p} = 2$ mm and $d = 1$ mm, $N^{\mathrm{DIR}} = 16$). Rather, it represents a minimum statistical requirement to ensure that the mean DIR registration error is an accurate representation of the spatial accuracy performance for any given case. Furthermore, $\bar{s}_{\mathrm{DIR}}^{p}$ is generally not known *a priori*, and therefore must be estimated from large samples of uniformly distributed error measurements. It is highly recommended that datasets that are made publicly available for purposes of multi-institutional comparative evaluation studies consist of sufficiently large validation test points to avoid drawing erroneous conclusions based on insufficient data. In the design of such studies, estimates of respective values for $\bar{s}_{\mathrm{DIR}}^{p}$ should be obtained for all n participating algorithms from prior evaluation studies, incorporating associated uncertainties, to calculate appropriate validation landmark sample sizes. For the specified interval range, d, the minimum allowable sample size used for the specific study should then be given by $\max\left\{N_{\pm d}^{\mathrm{DIR}_1}, \ldots, N_{\pm d}^{\mathrm{DIR}_n}\right\}$. We propose $\bar{s}_{\mathrm{DIR}}^{p}$, a characteristic of the given algorithm and the anatomic target, be measured on acceptance testing or comparative evaluation of new DIR algorithms. It is also noteworthy that the same sample size considerations apply regardless of whether the validation test points are delineated in patient or phantom images.

5. Discussion

Expert-determined landmark correspondences have become a widely adopted reference for evaluating DIR accuracy for lung image data; however, there has been great variability in their use. In this study we have presented a framework for objective evaluation of thoracic DIR spatial accuracy, based on the use of large samples of expert-determined landmark feature pairs between volumetric images as a reference for spatial accuracy measurements. A summary of the methodology is presented as follows.

5.1. Selection of anatomical landmark pairs

The use of registered landmarks as an objective metric for evaluation of image registration loses its significance if the point correspondences are calculated automatically. Thus, it is crucial that the individual feature points are first selected, and then manually registered between image volumes by a human observer, with expertise in imaging of the appropriate anatomic site. This is undoubtedly a lengthy task, and it is difficult to appreciate the necessity of enduring the process, without having some prior demonstration as to why there is a necessity. Thus, there has been no compelling reason for investigators to pursue what would presently be considered as unnecessarily large datasets. In fact, large datasets are crucial; thus, the process should be streamlined as best as possible, while still leaving the actual registration of individual feature points in the hands of the expert. Simple software design considerations can be highly effective in this regard.

The APRIL software utilized here was developed to maximize the number of landmark point pairs that a user can select. For any designated source landmark, an optional computer assistance tool provided rapid localization of an estimated target correspondence, based on user-determined threshold and search range criteria. However, the final selection of the corresponding point was performed manually by the user to ensure that the selection represents the expert choice and not the particular calculated estimate. A range of 1166–1561 unique anatomical features were manually identified and tracked between the five individual pairs of treatment planning CT images.

For summary statistics such as mean and standard deviation of the measured registration errors to accurately reflect the DIR performance throughout the lung, points must be distributed sufficiently uniformly in space, such that spatial variations in the DIR accuracy are detected. Perhaps the most significant perceived drawback regarding the use of manually registered feature points for objective evaluation of DIR is the notion that naturally occurring anatomical features are too few and too unevenly distributed to provide for rigorous performance evaluation. One reason for this view may be the requirement by some investigators that the anatomic identification of each landmark point is necessary (e.g. the *n*th generation of the right main bronchus), similar to landmark registration in neuroimaging applications. In contrast, we feel that the expert user must simply uniquely identify corresponding image features without identifying their exact anatomic location.

5.2. Characterization of landmark datasets

Estimates of variability within the primary reader (intra-observer) and among readers (inter-observer) for matching the corresponding landmark features must be obtained. In general, the variance sets a lower limit on the spatial accuracy that is detectable using the validation landmarks. This characterization process also requires large samples of measurements to ensure tight confidence intervals on the estimates of observer variance.

The complete reference displacement set consisting of five lung CT image pairs and 6762 landmark point pairs was statistically characterized with measurements of the intra- and inter-observer variance by repeated registration of multiple subsets of feature points. An important factor not specifically investigated in this study is the effect of an observer's experience or familiarity with the APRIL software on the manual registration process. This is an important point to consider because the ability to resolve registration errors is largely a function of how well the validation points can be reproduced. In practice, care should be taken to ensure adequate training in the manual registration process prior to the acquisition of formal repeated registration measurements. A more thorough characterization of the observer variance based on larger populations of participating observers is still necessary.

5.3. Evaluation of DIR

The sets of validation landmarks were utilized to perform quantitative comparative evaluation of a gradient-based OFM algorithm and a landmark interpolation algorithm based on MLS (Schaefer *et al* 2006). The validation landmarks provided for statistical tests on mean registration errors, as well as visual and quantitative assessment of spatial accuracy performance with location and magnitude displacement. It should be emphasized that the goal of this study was not to perform an explicit comparison between landmark-based MLS and gradient-based OFM for thoracic DIR. Though the OFM results presented here were indeed poor, further improvement based on optimization of internal parameters such as the regularization smoothing parameter (α) and the iterations of equation (2) could almost certainly be achieved. Furthermore, the spatial accuracy of the MLS DIR is a function of the quantity and uniformity of the input point pairs used for MLS interpolation. Thus, variations in input landmark selection will result in variable DIR output. Optimization of input parameters for both algorithms could be investigated further based on the evaluation methods presented in this study.

A great deal of information is provided by a large landmark set between even a single pair of volumetric images. As more patient cases become available, and as the validation feature points are propagated onto the remaining phases of the 4D CT datasets, a more complete and statistically sound characterization of DIR spatial accuracy performance can be achieved. The reference data will be invaluable for optimization of algorithms under development, as the error analysis procedure can be entirely automated to generate formatted error reports as part of the DIR output. This could largely streamline comparative evaluation studies and allow for more detailed ranking of multi-algorithm spatial accuracy performance that is based on more than simple summary error statistics. With these procedures in place, the problem of formal acceptance testing of a DIR algorithm can be posed as deciding which performance characteristics are most relevant for the given application, and whether or not the confidence intervals on the measured characteristic errors are acceptable. However, during this process, only those error measurements obtained using patient images equivalent (e.g. 4D CT) to those that will be encountered in clinical practice should be considered.

In order to properly interpret published reports of DIR spatial accuracy for which different reference datasets were utilized, it is important that landmark-based evaluation studies of DIR provide error measurements with clear indication of the observer variance and motion characteristics of the data points, as well as image resolution and voxel dimensions.

5.4. Minimum statistical requirements on sample size

Using numerical simulation, we demonstrated that the statistical uncertainty of the DIR spatial error estimate is inversely proportional to the square root of the number of landmark point pairs and directly proportional to the standard deviation of the spatial error specific to the DIR (SD_{DIR}) (figure 8). From these statistical considerations and from demonstration of the variation in spatial accuracy with displacement size and anatomic location, we propose that large (>1000) validation landmark sets are indeed necessary for rigorous evaluation of DIR spatial accuracy in the lung.

For comparative evaluation and/or validation of DIR, summary statistics such as mean registration error and standard deviation should comprise only a component of the overall characterization of DIR spatial accuracy performance, regardless of the sample size from which they are derived. More detailed analyses should be performed investigating the characterization of spatial accuracy with regard to clinically relevant variables that could potentially affect DIR output. This necessarily requires large sample validation datasets and multiple test cases to ensure meaningful statistics for the range of potential clinical variables. In prior studies, a maximum of 108 unique anatomical landmarks, divided between right and left lungs, have been manually identified within a single volumetric image pair for DIR performance assessment (Al-Mayah *et al* 2008). Sarrut *et al* have recently reported on a landmark set surpassing 500 feature points distributed over four 4D CT phases and three patients (Sarrut *et al* 2007). Analyses based on landmark samples that are not sufficient in size, or that are restricted to highly selective features, risk under-estimating the mean spatial error and SD_{DIR}. In this study, we have demonstrated that large (>1100) validation landmark datasets are indeed feasible for rigorous evaluation of DIR spatial accuracy in the lung. To our knowledge, the cases presented here represent the most extensive and comprehensive set of expert-determined landmark correspondences to date. Efforts are currently underway utilizing the APRIL software to construct a library of manually registered 4D CT datasets, with a requirement of >1000 unique landmarks per case, to facilitate comparative evaluation of DIR for thoracic CT. These data will be made publicly available through our website, http://www.dir-lab.com.

5.5. Application to QA

Currently, it is not clear that the selection of such large validation landmark sets could ultimately prove feasible for application to routine QA assessment of DIR. However, the presented framework for rigorous evaluation suggests that it may not be necessary either. Ideally, evaluation and characterization of the spatial accuracy performance of a given algorithm should be established prior to clinical acceptance. As mentioned above, those decisions could be based on some evaluation of the performance characteristics that are most relevant for the specific application and whether or not the measured characteristic errors are well defined and acceptable. Assuming that the characterization was based on multiple test cases, each of which consists of relatively large (e.g. >1000) validation landmark sets, one could derive estimates of the landmark sample size necessary to obtain confidence intervals of a specified length and statistical significance about the mean registration error. Thus, only a modest sample size may be necessary to obtain the desired summary error statistics for an arbitrary case. For example, for the two DIR algorithms tested in this study, the sample size requirements for 95% CIs of ±0.5 mm are 1050 (OFM) and 36 (MLS). The insight obtained in the prior validation process would be directly applicable for assessing the potential for regional registration errors not necessarily reflected by the sparse set of QA landmarks. In practice, it is likely that a combination QA strategy consisting of landmark point pairs and perhaps some combination of independent evaluations of global DIR performance will prove most effective. For example, Zhong *et al* (2007) have recently reported on a finite-element-based metric for assessing global DIR performance. In that study, the authors propose an automated method for detecting components of the calculated displacement fields that violate principles of continuum mechanics. The concept of unbalanced energy is introduced as an indicator for regions in which the DIR transformation is thought to be of poor quality. Though the proposed method does not provide for direct quantitative assessment of DIR spatial accuracy, it suggests a means for automatically delineating regions in which the DIR performance is suspect. Additional landmark pair locations could then be weighted toward those suspect regions for local quantitative evaluation. The combination of this type of global assessment with landmark-based measurements of spatial accuracy may prove to be an effective and practical strategy for QA of DIR on a routine clinical basis.

6. Conclusion

We have presented a framework and corresponding software infrastructure for rigorous quantitative evaluation of DIR spatial accuracy. The feasibility of generating large (>1100) validation landmark sets has been demonstrated on five component phase pairs from clinically acquired treatment planning 4D CT data. The results demonstrate that large landmark point sets provide an effective means for objective evaluation of DIR with a narrow uncertainty range, and suggest a practical strategy for QA of DIR spatial accuracy on a routine clinical basis.

Acknowledgments

We most sincerely thank the University of Texas M. D. Anderson Cancer Center Physician–Scientist Program that provided support for this project. This work was also partially supported by the National Cancer Institute through NIH/NCI Grant (R21CA128230) and through an NIH Training Grant (T32CA119930). This work was also supported in part by The Gulf Coast Center for Computational Cancer Research with the GC4R Seed Grant for Collaborative Advances in Biomedical Computing (C-ABC).

References

Al-Mayah A, Moseley J and Brock K K 2008 Contact surface and material nonlinearity modeling of human lungs *Phys. Med. Biol.* **53** 305–17

Beauchemin S S and Barron J L 1995 The computation of optical flow *ACM Comput. Surv.* **27** 433–66

Bevington P R and Robinson D K 2003 *Data Reduction and Error Analysis* (New York: McGraw-Hill) vol 3

Boldea V, Sharp G C, Jiang S B and Sarrut D 2008 4D-CT lung motion estimation with deformable registration: quantification of motion nonlinearity and hysteresis *Med. Phys.* **35** 1008–18

Brock K K, Nichol A M, Menard C, Moseley J L, Warde P R, Catton C N and Jaffray D A 2008 Accuracy and sensitivity of finite element model-based deformable registration of the prostate *Med. Phys.* **35** 4019–25

Brock K K, Sharpe M B, Dawson L A, Kim S M and Jaffray D A 2005 Accuracy of finite element model-based multi-organ deformable image registration *Med. Phys.* **32** 1647–59

Casella G and Berger R 2001 *Statistical Inference* (New York: Duxbury Press)

Coselmon M M, Balter J M, McShan D L and Kessler M L 2004 Mutual information based CT registration of the lung at exhale and inhale breathing states using thin-plate splines *Med. Phys.* **31** 2942–8

Dawood M, Buther F, Jiang X and Schafers K P 2008 Respiratory motion correction in 3-D PET data with advanced optical flow algorithms *IEEE Trans. Med. Imaging* **27** 1164–75

Fitzpatrick J M 2001 *Medical Image Registration* ed M R Neuman (New York: CRC Press)

Gee J C 2000 Performance evaluation of medical image processing algorithms *Medical Imaging 2000: Image Processing* (San Diego, CA: SPIE) pp 19–27

Gonzalez R C and Woods R E 2008 *Digital Image Processing* (Upper Saddle River, NJ: Prentice-Hall)

Guerrero T, Sanders K, Castillo E, Zhang Y, Bidaut L, Pan T and Komaki R 2006 Dynamic ventilation imaging from four-dimensional computed tomography *Phys. Med. Biol.* **51** 777–91

Guerrero T, Zhang G, Huang T C and Lin K P 2004 Intrathoracic tumour motion estimation from CT imaging using the 3D optical flow method *Phys. Med. Biol.* **49** 4147–61

Horn B K P and Schunck B G 1981 Determining optical flow *Artif. Intell.* **17** 185–203

Kaus M R, Brock K K, Pekar V, Dawson L A, Nichol A M and Jaffray D A 2007 Assessment of a model-based deformable image registration approach for radiation therapy planning *Int. J. Radiat. Oncol. Biol. Phys.* **68** 572–80

Keall P J, Joshi S, Vedam S S, Siebers J V, Kini V R and Mohan R 2005 Four-dimensional radiotherapy planning for DMLC-based respiratory motion tracking *Med. Phys.* **32** 942–51

Lehmann T M 2002 From plastic to gold: a unified classification scheme for reference standards in medical image processing *SPIE Medical Imaging* ed M Sonka and J M Fitzpatrick (San Diego, CA: Society of Photo-Optical Instrumentation Engineers) pp 1819–27

Lewis J P 1995 Fast template matching *Vis. Interface* 120–3

Li P, Malsch U and Bendl R 2008 Combination of intensity-based image registration with 3D simulation in radiation therapy *Phys. Med. Biol.* **53** 4621–37

Lu W, Chen M L, Olivera G H, Ruchala K J and Mackie T R 2004 Fast free-form deformable registration via calculus of variations *Phys. Med. Biol.* **49** 3067–87

Lu W, Olivera G H, Chen Q, Ruchala K J, Haimerl J, Meeks S L, Langen K M and Kupelian P A 2006 Deformable registration of the planning image (kVCT) and the daily images (MVCT) for adaptive radiation therapy *Phys. Med. Biol.* **51** 4357–74

Pevsner A *et al* 2006 Evaluation of an automated deformable image matching method for quantifying lung motion in respiration-correlated CT images *Med. Phys.* **33** 369–76

Press W H, Teukolsky S A, Vetterling W T and Flannery B P 2002 *Numerical Recipes in C++: The Art of Scientific Computing* (Cambridge: Cambridge University Press)

Rietzel E and Chen G T Y 2006 Deformable registration of 4D computed tomography data *Med. Phys.* **33** 4423–30

Sarrut D, Boldea V, Miguet S and Ginestet C 2006 Simulation of four-dimensional CT images from deformable registration between inhale and exhale breath-hold CT scans *Med. Phys.* **33** 605–17

Sarrut D, Delhay S, Villard P F, Boldea V, Beuve M and Clarysse P 2007 A comparison framework for breathing motion estimation methods from 4-D imaging *IEEE Trans. Med. Imaging* **26** 1636–48

Schaefer S, McPhail T and Warren J 2006 Image deformation using moving least squares *ACM SIGGRAPH 2006 Papers* (Boston, MA: ACM Press)

Sharpe M and Brock K K 2008 Quality assurance of serial 3D image registration, fusion, and segmentation *Int. J. Radiat. Oncol. Biol. Phys.* **71** S33–7

Wang H, Dong L, Lu M F, Lee A L, de Crevoisier R, Mohan R, Cox J D, Kuban D A and Cheung R 2005a Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy *Int. J. Radiat. Oncol. Biol. Phys.* **61** 725–35

Wang H, Dong L, O'Daniel J, Mohan R, Garden A S, Ang K K, Kuban D A, Bonnen M, Change J Y and Cheung R 2005b Validation of an accelerated demons algorithm for deformable image registration in radiation therapy *Phys. Med. Biol.* **50** 2887

Wolthaus J W H, Sonke J J, van Herk M and Damen M F 2008 Reconstruction of a time-averaged midposition CT scan for radiotherapy planning of lung cancer patients using deformable registration *Med. Phys.* **35** 3998–4011

Wu Z, Rietzel E, Boldea V, Sarrut D and Sharp G C 2008 Evaluation of deformable registration of patient lung 4D CT with subanatomical region segmentations *Med. Phys.* **35** 775–81

Zhong H, Peters T and Siebers J V 2007 FEM-based evaluation of deformable image registration for radiation therapy *Phys. Med. Biol.* **52** 4721–38