

Evaluation of 4D-CT Lung Registration

Sven Kabus¹, Tobias Klinder^{1,2}, Keelin Murphy³, Bram van Ginneken³,
Cristian Lorenz¹, and Josien P.W. Pluim³

¹ Philips Research Europe – Hamburg, Germany
sven.kabus@philips.com

² Institut für Informationsverarbeitung, Leibniz University of Hannover, Germany

³ University Medical Center, Utrecht, The Netherlands

Abstract. Non-rigid registration accuracy assessment is typically performed by evaluating the target registration error at manually placed landmarks. For 4D-CT lung data, we compare two sets of landmark distributions: a smaller set primarily defined on vessel bifurcations as commonly described in the literature and a larger set being well-distributed throughout the lung volume. For six different registration schemes (three in-house schemes and three schemes frequently used by the community) the landmark error is evaluated and found to depend significantly on the distribution of the landmarks. In particular, lung regions near to the pleura show a target registration error three times larger than near-mediastinal regions. While the inter-method variability on the landmark positions is rather small, the methods show discriminating differences with respect to consistency and local volume change. In conclusion, both a well-distributed set of landmarks and a deformation vector field analysis are necessary for reliable non-rigid registration accuracy assessment.

1 Introduction

Image registration plays an indispensable role in medical application areas such as diagnosis, therapy planning, and follow-up assessment. Evaluating its accuracy has therefore become an issue of growing importance [1,2,3]. Contrary to rigid registration methods, validation of non-rigid methods remains a challenging task due to the absence of a gold standard [3]. Currently, this task is addressed by computing the target registration error (TRE) based on anatomical landmarks or on surfaces (e.g., contoured masses) in the images to be registered. The resulting TRE is widely accepted as an indicator for the method's accuracy. TRE-based evaluation for lung applications, however, suffers from low contrast in near-to-pleura regions, in particular if the scans are acquired dynamically as in 4D-CT. Consequently, evaluation is typically performed on a limited set of landmarks defined on major bifurcations of the vessel tree (cf., e.g., [4,1,5]), often grouped around the mediastinum and thus not covering the lungs entirely. Since registration accuracy in near-mediastinal regions may not be representative for, e.g., the lower lungs as here the motion amplitude is largest, it is therefore of particular interest how the choice of landmarks affects the estimated overall registration accuracy.

On the basis of a publicly available and carefully validated 4D-CT chest dataset (POPI-model [4]) we compare in this work the influence of two different landmark sets on the TRE: (a) a set consisting of 37 lung landmarks grouped around the mediastinum (obtained by the POPI-model), (b) a set of 100 landmarks being well-distributed throughout the lung volume (generated by a semi-automatic tool [6]). For the registration accuracy, six different registration algorithms are compared in terms of the TRE (per landmark set and per lung region) as well as the inter-method variability at the landmark positions. In addition, an analysis of the resulting deformation vector fields is included.

2 Methods

2.1 Datasets

The basis of the evaluation is a respiratory gated chest CT dataset (10 phases, $0.98 \times 0.98 \times 2.00 \text{mm}^3$ voxel size) provided by the 'POPI-model' [4]. Maximum diaphragm motion is measured as about 18mm (9mm) for the right (left) lung. In addition, the POPI-model provides a set of 41 landmarks annotated in each breathing phase. From these, 37 are located within the lungs and are employed for evaluation (*landmark reference set (LRS) A*). Moreover, an automatic tool [6] is used to create a set of 100 well-distributed lung landmarks in a reference phase. These landmarks are then semi-automatically propagated [6] to all other respiratory phases individually by two trained scientists (inter-observer difference of 1.0mm); their averaged annotations finally defines the second 4D set of landmark positions (*LRS B*). For a region-based analysis, each lung is divided into four volumetric regions such that the near-mediastinal parenchyma is enclosed in a sphere around the lungs' center of mass (regions I and II) and the remaining parenchyma of each lung is separated into an upper, central, and lower region. The dispersal of landmarks concerning the defined regions is shown in Fig. 1 for both landmark reference sets. While the landmarks from set A mainly belong to regions I and II, the landmarks from set B show a more regular distribution.

2.2 Registration Schemes

This work compares six fully automatic registration schemes, each computing deformation vector fields (DVF) pointing from the end-inhale phase onto all other respiratory phases. Three methods are in-house, three other methods have either been downloaded or have been already applied to the data set. Five of the six methods are volumetric schemes whereas one is surface-based. Among the volumetric schemes, three are based on a parametric B-spline representation, while the two others are non-parametric.

Surface-based Registration (MBS). [7] After extracting lung surface and vessel tree a method of deformable surface models is applied on the generated

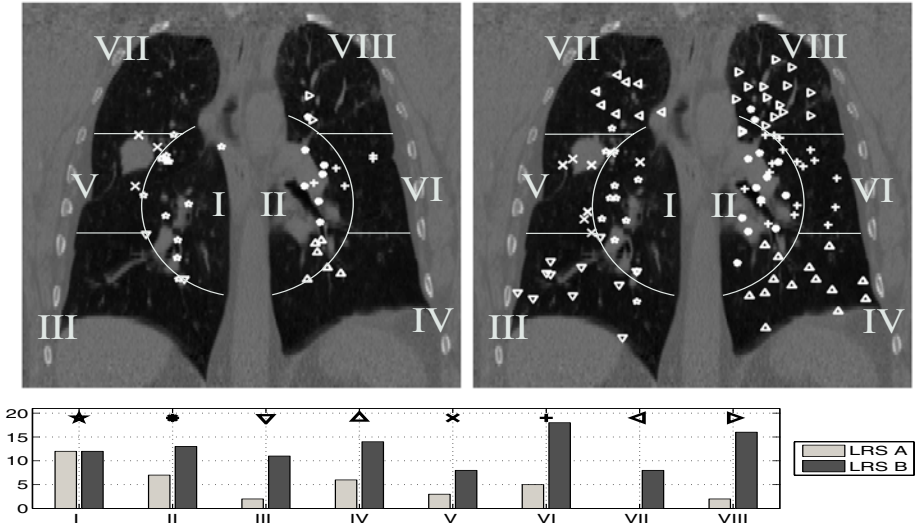


Fig. 1. Region definition and landmark positions shown in a coronal projection for the landmark reference set A (top left) and for the landmark reference set B (top right) together with the dispersal of landmarks among the regions (bottom). For improved visualization each region is given a unique landmark symbol.

Table 1. Property overview for the volumetric registration schemes. All schemes are used in a multi-level embedding.

Method	transformation class	cost function	regularizer	use of lung mask
POPI-par [8]	B-splines	SSD	–	no
elastix [9]	B-splines	NCC	–	yes
DROP [10]	B-splines	SAD	grid vector distances	yes
POPI-nonpar [5]	non-par.	SSD	Gaussian + linear elastic	no
FEIR [11]	non-par.	SSD	linear elastic	yes

iso-surface to propagate the mesh to all other phases by minimizing an energy term. A dense DVF is then obtained using a thin-plate-spline interpolation scheme.

Volumetric Registrations [5,8,9,10,11]. An overview of the employed volumetric schemes is given in Tab. 1, for further details we refer to the cited references. POPI-par is a modification based on a frequently used FFD-approach [12] whereas POPI-nonpar describes a modified Demon’s approach [5].

MBS, elastix and FEIR are in-house methods and are applied with standard parameter settings. DROP is applied following the published instructions [10]. The computation for these methods was completed after 3–15 minutes per phase on standard PCs. The DVFs for POPI-par and POPI-nonpar are not computed by the authors themselves, these are already provided by the POPI-model.

3 Results

To ensure the correct use of the output of the different registration schemes, each DVF is applied to the reference phase to carefully check the correct alignment of the pleura with the corresponding phase.

Then, for all the six methods each set of landmarks annotated in the reference phase is transformed to the other phases. By comparing the transformed position with the respective annotated position, the target registration error (TRE) is obtained for each landmark, respiratory phase, and registration method. From this, we deduce both method-based and LRS-based mean TRE and standard deviation by collecting the TREs from all phases and landmarks, cf. Tab. 2. In addition, the mean of the upper 10% quantile (MU10) is calculated, which, compared to the maximum, is less dominated by outliers. An inspection of Tab. 2 reveals a clear decrease of the TRE for all methods with elastix performing slightly better and POPI-nonpar slightly worse than the others. Focusing on the inter-LRS differences, we further notice that the MU10 values show a larger relative difference than the mean values. The reason can be found in lung regions covered by LRS B but not by LRS A (cf. the landmark distribution displayed in Fig. 1) – on the one hand this refers to regions located in the apex of the lungs with a quite small TRE, on the other hand to regions in the lower lungs with a larger TRE (demonstrated in Fig. 2). Particularly in the lower right lung (region III) the mean-TRE (2.4–3.1mm) as well as the MU10-TRE (6.1–8.7mm) are higher by a factor three (cf. Tab. 3) probably caused by the larger motion amplitude in the right lung. However, since Fig. 1 reports only 2 landmarks in region III for LRS A, this inhomogeneous distribution of the TRE cannot be deduced from a landmark distribution grouped around the mediastinum.

As the TRE differs only slightly among the registration methods, we investigate the inter-method variability evaluated on the displaced landmark positions of both landmark sets. To this end, for each landmark position and each phase individually, the sum of eigenvalues of the covariance matrix is computed from the displacement vectors of all methods. For an LRS-based discrimination, its mean is taken over all landmarks and phases yielding a value of 0.31mm for set A and one of 0.47mm for set B, thus an increased inter-method variability

Table 2. TRE for each method and LRS (shown are mean, standard deviation, and MU10 in mm)

Method	Landmark reference set A	Landmark reference set B
without registration	3.68 ± 2.97 (9.6)	4.22 ± 3.20 (11.2)
MBS	1.11 ± 0.65 (2.5)	1.36 ± 1.13 (3.9)
POPI-par	1.07 ± 0.56 (2.2)	1.26 ± 1.17 (3.8)
elastix	0.96 ± 0.56 (2.1)	1.16 ± 0.97 (3.4)
DROP	0.98 ± 0.56 (2.2)	1.25 ± 1.07 (3.6)
POPI-nonpar	1.28 ± 0.42 (2.1)	1.41 ± 1.22 (4.1)
FEIR	1.05 ± 0.57 (2.2)	1.27 ± 1.00 (3.6)

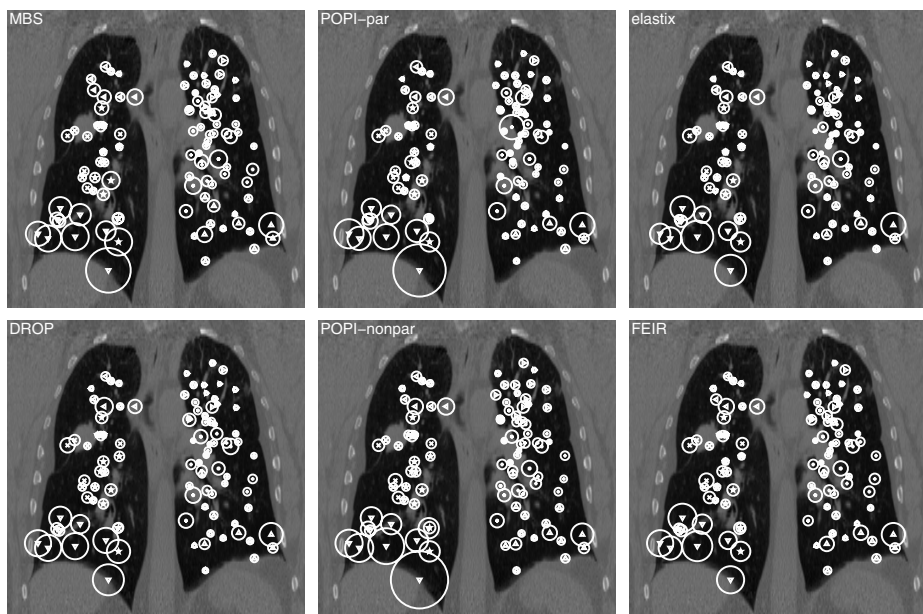


Fig. 2. TRE averaged over all respiratory phases for landmark reference set B shown proportional to the spheres' diameters

Table 3. TRE per lung region (cf. Fig. 1) for landmark reference set B (shown are mean, standard deviation, and MU10 in mm)

Method	Region I	Region II	Region III	Region IV
without reg.	5.28 ± 3.58 (11.5)	3.65 ± 2.00 (6.9)	7.31 ± 5.25 (15.7)	4.83 ± 2.77 (9.1)
MBS	1.44 ± 1.10 (3.7)	1.25 ± 0.73 (2.7)	2.69 ± 2.25 (7.4)	1.30 ± 0.87 (3.0)
POPI-par	1.21 ± 0.77 (2.6)	1.28 ± 0.96 (3.2)	2.71 ± 2.47 (7.9)	1.12 ± 0.86 (2.8)
elastix	1.17 ± 0.79 (2.8)	1.09 ± 0.69 (2.4)	2.37 ± 1.89 (6.1)	1.09 ± 0.80 (2.6)
DROP	1.33 ± 0.89 (3.1)	1.20 ± 0.75 (2.7)	2.49 ± 2.17 (7.4)	1.11 ± 0.80 (2.7)
POPI-nonpar	1.38 ± 0.77 (2.8)	1.22 ± 0.66 (2.5)	3.14 ± 2.60 (8.7)	1.25 ± 0.77 (2.6)
FEIR	1.21 ± 0.82 (2.9)	1.08 ± 0.66 (2.3)	2.48 ± 1.83 (6.2)	1.18 ± 0.86 (2.8)
Method	Region V	Region VI	Region VII	Region VIII
without reg.	4.64 ± 3.12 (10.4)	3.64 ± 1.94 (6.8)	3.58 ± 2.03 (7.1)	1.99 ± 1.15 (4.2)
MBS	1.24 ± 0.68 (2.7)	0.90 ± 0.46 (1.9)	1.12 ± 0.57 (2.3)	0.89 ± 0.47 (1.9)
POPI-par	1.15 ± 0.72 (2.7)	0.92 ± 0.50 (2.0)	1.11 ± 0.55 (2.2)	0.73 ± 0.39 (1.5)
elastix	1.27 ± 0.65 (2.6)	1.01 ± 0.57 (2.2)	1.14 ± 0.66 (2.5)	0.82 ± 0.45 (1.8)
DROP	1.19 ± 0.62 (2.4)	1.12 ± 0.62 (2.5)	1.30 ± 0.60 (2.4)	0.90 ± 0.41 (1.7)
POPI-nonpar	1.29 ± 0.73 (2.9)	1.10 ± 0.53 (2.2)	1.28 ± 0.66 (2.6)	1.01 ± 0.44 (1.8)
FEIR	1.41 ± 0.83 (3.1)	1.10 ± 0.62 (2.5)	1.17 ± 0.64 (2.4)	0.86 ± 0.44 (1.7)

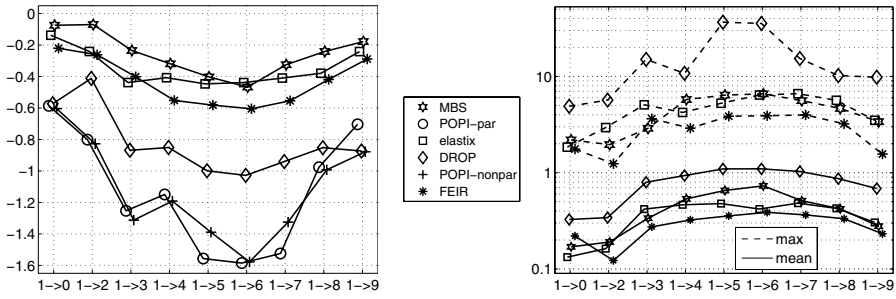


Fig. 3. Left: Minimum volume change (obtained from $\det(\text{Jac}(\text{DVF})) - 1$) over all lung voxels shown for each method and phase transition (values smaller than -1 indicate folding of the DVF); right: Mean and maximum consistency error (in mm) over all lung voxels shown for each phase transition

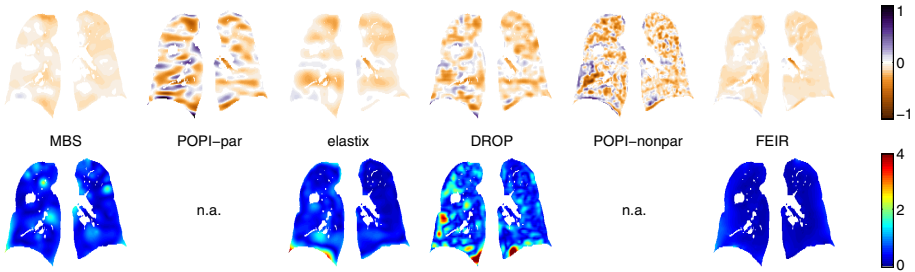


Fig. 4. For an exemplary coronal slice (same as in Fig. 1), local volume change (top) and local consistency error (bottom) are displayed after registration from end-inhale onto end-exhale. Note that for the local volume change maps blue (brown) colors indicate expansion (contraction).

on the well-distributed landmark set. Discriminating on a region-by-region basis determines a value of 1.81mm for region III and a range of 0.24–0.36mm for the other regions, thus a high variability for regions with a large TRE and vice versa. Finally, from a computation of the Mahalanobis distance, elastix and FEIR are identified to deviate less from the other methods (mean values of 2.21 and 2.09, respectively) whereas POPI-nonpar shows the largest deviation (6.66).

However, evaluation so far is restricted to the sparse sets of distinct landmark positions. To further investigate the inter-method variability and also to achieve greater insight into the methods’ properties an analysis of the DVFs is performed.

At first, we compute the determinant of the Jacobian of the DVF to measure the local change in volume which allows for detecting non-bijectively mapped regions (i.e., ‘folding’ occurs). Since each DVF is defined in the end-inhale domain we expect a contraction in general. An inspection of the minimum values (within the lungs only) revealed large differences: whereas MBS, elastix and FEIR show relatively little local contraction values, DROP and the two POPI-methods result in heterogeneous contraction-expansion-patterns (cf. Fig. 4, top) with severe foldings of the POPI-methods for phases 3 to 7 (cf. Fig. 3, left).

Secondly, for registration methods with both forward and backward DVFs available, we compute a local consistency measure. The consistency is determined by composing the forward and the backward DVFs and measuring the amplitude of the geometrical discrepancy at each position. Since none of the methods under consideration is consistent by definition, measuring the consistency error is a suitable indicator of how independent the registration result is from the image input order. Here, the smallest error is achieved by FEIR, the largest by DROP (cf. Fig. 3, right, and Fig. 4, bottom). Neither POPI-method is evaluated since backward DVFs were required.

4 Discussion and Conclusion

On the basis of a publicly available 4D-CT chest dataset we have investigated the impact of two different landmark distributions on non-rigid registration accuracy assessment: the first set of landmarks consists of 37 lung annotations grouped around the mediastinum as commonly described in the literature, the second set provides 100 annotations being well-distributed throughout the lung volume. While the mean target registration error on both landmark sets differs only slightly, a region-based analysis reveals smaller errors in apical regions but also a significantly higher error in the lower right lung and therefore a dependence of the target registration error on the distribution of landmarks. This dependence is observed for each registration scheme chosen from a collection of six popular methods including surface- and volume-based as well as parametric and non-parametric methods. Clearly, validation on further patient datasets is required, however, since generally the lower lungs show the largest motion amplitude, we believe that our result is of general validity and that a well-distributed set of landmarks is a necessity for reliable registration accuracy assessment.

Focusing on the inter-method variability we observed similar displacement vectors at the landmark positions. The methods can, however, be discriminated by a closer analysis of the resulting deformation vector fields, e.g., regarding the local change in volume. This parameter is clinically relevant since it is directly related to local lung ventilation estimation and has impact on lung diagnosis as well as on radiotherapy planning. Inspecting the local volume change discloses for three of the six methods several regions where the vector field is near to or already suffers from folding. Moreover, the methods differ in showing individual contraction-expansion-patterns. With the lung at end-inhale as the reference position in mind, an overall contraction – rather than a mixture of contracting and expanding regions – seems to be desirable from a physiological point of view. Another indication for inter-method variability is obtained from a local consistency analysis. Although non-consistent by definition, three out of four methods result in nearly consistent deformation vector fields and are thus independent of the image input order. Since similar target registration errors does not mean that different registration methods give similar results overall, an analysis of the deformation vector field is necessary to evaluate non-rigid registration accuracy.

Acknowledgments. We thank Heike Ruppertshofen for carrying out the landmark propagation as one observer. The POPI-dataset was acquired from the Léon Bérard Cancer Center, Lyon, France.

References

1. Brock, K.K.: On Behalf of the Deformable Registration Accuracy Consortium: A Multi-Institution Deformable Registration Accuracy Study. In: I. J. Radiation Oncology, ASTRO 2007, vol. 69, p. S44 (2007)
2. Vik, T., Kabus, S., von Berg, J., Ens, K., Dries, S., Klinder, T., Lorenz, C.: Validation and comparison of registration methods for free-breathing 4D lung CT. In: Proc. of SPIE Medical Imaging, vol. 6914, 69142P-1-69142P-10 (2008)
3. Christensen, G.E., Geng, X., Kuhl, J.G., Bruss, J., Grabowski, T.J., Pirwani, I.A., Vannier, M.W., Allen, J.S., Damasio, H.: Introduction to the non-rigid image registration evaluation project (NIREP). In: Pluim, J.P.W., Likar, B., Gerritsen, F.A. (eds.) WBIR 2006. LNCS, vol. 4057, pp. 128-135. Springer, Heidelberg (2006)
4. Vandemeulebroucke, J., Sarrut, D., Clarysse, P.: The POPI-model, a point-validated pixel-based breathing thorax model. In: Proc. of ICCR (2007)
5. Sarrut, D., Boldea, V., Miguet, S., Ginestet, C.: Simulation of 4D CT images from deformable registration between inhale and exhale breath-hold CT. *Med. Phys.* 33(3), 605-617 (2006)
6. Murphy, K., van Ginneken, B., Pluim, J.P.W., Klein, S., Staring, M.: Semi-automatic reference standard construction for quantitative evaluation of lung CT registration. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5242, pp. 1006-1013. Springer, Heidelberg (2008)
7. von Berg, J., Barschdorf, H., Blaffert, T., Kabus, S., Lorenz, C.: Surface based cardiac and respiratory motion extraction for pulmonary structures from multi-phase CT. In: Proc. of SPIE Medical Imaging, vol. 6511, 65110Y (2007)
8. Delhay, B., Clarysse, P., Pera, C., Magnin, I.E.: A spatio-temporal deformation model for dense motion estimation in periodic cardiac image sequences. In: Proc. of the workshop From Statistical Atlases to Personalized Models: Understanding Complex Diseases in Populations and Individuals, MICCAI, pp. 87-90 (2006)
9. Klein, S., Staring, M., Pluim, J.P.W.: Evaluation of optimization methods for non-rigid medical image registration using mutual information and B-splines. *IEEE Trans. Image Proc.* 16(12), 2879-2890 (2007)
10. Glocker, B., Komodakis, N., Tziritas, G., Navab, N., Paragios, N.: Dense image registration through MRFs and efficient linear programming. *Med. Image Analysis* 12(6), 731-741 (2008)
11. Kabus, S., von Berg, J., Yamamoto, T., Opfer, R., Keall, P.J.: Lung ventilation estimation based on 4D-CT imaging. In: Proc. of the first international workshop on pulmonary image analysis, MICCAI, pp. 73-82 (2008)
12. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Trans. Med. Imaging* 18(8), 712-721 (1999)