Semi-automatic Reference Standard Construction for Quantitative Evaluation of Lung CT Registration

K. Murphy, B. van Ginneken, J.P.W. Pluim, S. Klein, and M. Staring

University Medical Center, Utrecht, The Netherlands

Abstract. An algorithm is presented for the efficient semi-automatic construction of a detailed reference standard for registration in thoracic CT. A well-distributed set of 100 landmarks is detected fully automatically in one scan of a pair to be registered. Using a custom-designed interface, observers locate corresponding anatomic locations in the second scan. The manual annotations are used to learn the relationship between the scans and after approximately twenty manual marks the remaining points are matched automatically. Inter-observer differences demonstrate the accuracy of the matching and the applicability of the reference standard is demonstrated on two different sets of registration results over 19 CT scan pairs.

1 Introduction

The accurate registration of intra-patient thoracic CT scans has a variety of motivating clinical applications including improved ease of visual comparison, quantitative or automatic analysis of pathology progression, and in the case of inspiration/expiration pairs, analysis of lung function. In radiotherapy planning, registration information can be used to construct pulmonary motion models in order to propagate the location of the target region [7].

Although many promising registration algorithms exist, the quantitative evaluation of these techniques poses a further challenge due to the lack of an established reference standard. Urschler et al.[11] propose investigating performance by synthetically warping data such that the original image and the transformed image are known in advance as well as the ideal transform between them. This approach provides only a generic evaluation however and algorithm performance on real clinical data cannot be measured in this way. Other authors measure registration accuracy based on tumour overlap [12], nodule positions [2,1], or small numbers of manually annotated landmark positions [12] all of which provide information about the registration quality at only a very small and/or manually selected number of locations.

In this work a method is presented to formulate a registration reference standard for CT image pairs in an efficient semi-automatic manner resulting in a well-distributed mesh of corresponding landmarks throughout the lung volume in each image. The utility of this reference standard is demonstrated in the evaluation of a parametric intensity-based image registration method implemented

D. Metaxas et al. (Eds.): MICCAI 2008, Part II, LNCS 5242, pp. 1006–1013, 2008.

firstly with lung masks and secondly without any masking of the lung volume. Quantitative evaluation based on landmark correspondence enables confirmation that there is a significant difference between the results of the two registration techniques.

2 Materials

All scans used in this work form part of an experimental lung cancer screening programme. Nineteen patients (17 male, 2 female, ages 51-68yrs), each with a baseline and a follow-up scan (3-9 months apart) were chosen randomly from the database. All scans were obtained at full inspiration and without contrast injection on a 16 detector-row scanner (Mx8000 IDT or Brilliance 16P, Philips Medical Systems). They have a per-slice resolution of 512×512, with the number of slices per scan varying from 383 to 529. Slice thickness is 1mm with slice-spacing of 0.7mm. Pixel spacing in the X and Y directions varies from 0.55mm to 0.8mm.

3 Methods

3.1 Automatic Landmark Detection

The initial step in setting the reference standard is to automatically determine a number of landmark locations in the baseline scan for each patient. The landmarks are required to be well-distributed throughout the lung volume and to be identifiable on the corresponding follow-up scan.

The algorithm to automatically choose landmarks in the baseline scan, which is partially based on the work of Likar et al. [6], proceeds as follows: Points outside the lung volume are excluded from consideration. Within the lung volume, only every 5th point in each direction is considered in order to improve computational efficiency. Points on the pleural surface are also excluded since it is difficult to reliably match these anatomical locations in the follow-up scan.

For all remaining points $p(x, y, z)$ with intensity $I(x, y, z)$ a distinctiveness value $D(p)$ will be calculated estimating the dissimilarity of p with its surrounding region. Firstly, an estimate of the gradient value $G(p)$ is calculated by

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

where $G_x(p)$,$G_y(p)$ and $G_z(p)$ are directional gradients based on finite differences. Points where $G(p)$ is below the threshold $T_G = 300$ are excluded from further processing as they are likely to be extremely difficult to match reliably in the follow-up image. T_G was established experimentally during development.

Around each point p a hypothetical spherical surface with a radius of 8 voxels is constructed and 45 points, $q_1 ... q_{45}$, uniformly distributed on the surface are selected using the technique of Saff et al. [9]. A region of interest $ROI(q_i)$ around each point q_i is compared with the corresponding region of interest $ROI(p)$

around the original point p. $ROI(p)$ is defined as a spherical kernel of voxels centred at p with a radius of 5 voxels. The difference $\text{Diff}(ROI(p), ROI(q_i))$ is defined as the average absolute difference of the corresponding voxel intensities in the two ROIs. The distinctiveness value $D(p)$ is calculated for each point p as follows:

$$D(p) = \frac{G(p)}{max_j(G(p_j))} \sum_{i=1}^{45} \frac{\text{Diff}(ROI(p), ROI(q_i))}{45}$$

where j is the total number of points for which we calculate $D(p)$ in this scan. Approximately 1500 points per baseline scan are labelled with a distinctiveness value in this way. In addition to choosing the most distinctive points as landmarks we require an even distribution of the landmarks throughout the lungs. The points are therefore ordered with the most distinctive points first and chosen as follows:

1. The most distinctive point available is selected as a landmark as long as it is at least *minDist* voxels in distance from every other point selected so far.
2. When no more points meet this requirement, set *minDist*=*minDist*-10 and repeat step 1.
3. Continue until n landmarks have been selected.

In this study we set n=100 and an initial value of *minDist* = 400. A projection view of all the landmarks selected for a scan is shown in figure 1(a) while figure 1(b) shows some examples of landmark locations.

3.2 Establishing Landmark Correspondence

A semi-automatic system was developed to accurately match the voxels identified as landmarks in the baseline scan with voxels at the corresponding anatomic locations in the follow-up scan. Each scan pair was processed twice by independent observers (medical students). The observers were required to match at least 20

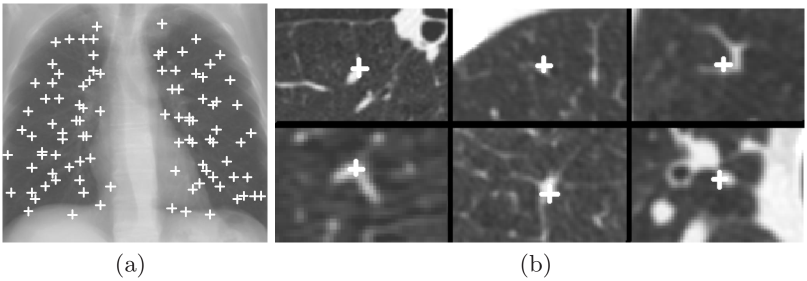
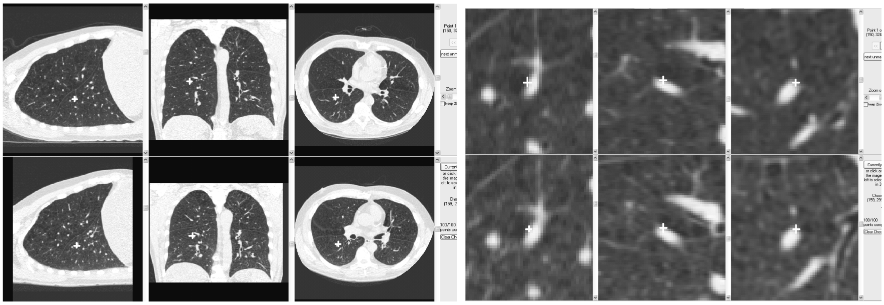


Fig. 1. (a)A set of automatically determined landmarks projected in the coronal direction. (b) Example landmarks shown at various zoom levels. Marker sizes have been increased for visualisation in these images.



(a) (b)

Fig. 2. The graphical user interface used to match points in a baseline scan (top row images) and a follow-up scan (bottom row images). (a)Zoomed out view. (b)Zoomed more closely on the landmark. Marker sizes have been increased for visualisation in these images.

of the 100 landmarks manually using a custom-made graphical interface. Subsequently, subject to certain conditions, the system matched the remaining points automatically. These two steps are described in more detail below. The annotation procedure took 20-30 minutes per scan-pair and did not require observers with significant experience of pulmonary anatomy.

Graphical User Interface. The system was designed to allow the observer to view the landmark l in question on the baseline scan in all three slice directions at one time. The follow-up scan was presented below in the same fashion. Screenshots from the system are shown in figure 2. The user could manually select the matching landmark location lm_{man} either by clicking on any point in one of the three views of the follow-up scan, or by scrolling through all three views individually until the three most appropriate slices were located. The observers were encouraged to view the presented and chosen landmarks at various zoom-levels and to confirm their final choice at the highest zoom level where individual voxels were clearly visible. They were permitted to repeatedly re-locate their matched landmark until they were satisfied with their choice.

Automatic Landmark Matching. The matching pairs of landmark correspondences manually annotated by the observer are used in the formation of a thin-plate-spline [3] warping of the follow-up image. This warping is not displayed to the user but used internally to represent the relationship between the baseline and follow-up images. Each point-pair (l, lm_{man}) manually annotated by the observer is added to the thin-plate-spline progressively improving its accuracy. When a new landmark point is presented to the observer the system utilizes the warped image to estimate where the anatomic match will be located in the follow-up scan. The location lm_{est} of the estimated match is used to determine which slices from the follow-up scan should be displayed initially to the observer. Thus, as the warped image becomes more accurate, the task of the observer becomes easier, with the initially displayed slices providing increasingly accurate starting points.

After some time the warped image is sufficiently accurate to enable the system to proceed with matching the remaining landmarks without user interaction. Automatic matching begins when a)The observer has manually matched at least 20 landmarks l with corresponding locations lm_{man} and b)The system has estimated 5 or more consecutive matches lm_{est} in a location within $\sqrt{6}$ voxels of the location lm_{man} indicated by the observer.

During the automatic matching the system improves on the location lm_{est} provided by the warped image by means of a local search around lm_{est} for a point whose surrounding region appears most similar to that of the landmark point l. All points p_i within 4 voxels of lm_{est} are considered as candidates. Cubic regions of interest $\mathrm{ROI}(l)$ and $\mathrm{ROI}(p_i)$ with sides of length 13 voxels are defined around the landmark point l and the point p_i under investigation. The p_i where the sum of squared differences between intensities in the regions of interest, $\mathrm{SSD}(\mathrm{ROI}(l), \mathrm{ROI}(p_i))$ is minimal is selected as the final automatic landmark match lm_{auto}. If a p_i is not found such that $\mathrm{SSD}(\mathrm{ROI}(l), \mathrm{ROI}(p_i)) < T_{SSD}$ then the match lm_{auto} is considered uncertain and the landmark is returned unmatched to the observer with the best system estimate as a suggestion. The threshold value T_{SSD} which was determined empirically during system development corresponds to a root mean square difference of approximately 213HU per voxel. This threshold was exceeded on approximately 2 landmarks per scan-pair and in most cases the system suggestion was very accurate allowing the observer to manually select the match lm_{man} without difficulty.

3.3 Registration Methods

Prior to registration the baseline and follow-up scans were down-sampled in order to reduce memory consumption. The calculated transform from the registration procedure was subsequently applied to the full resolution follow-up scan. The registration procedure consisted of an initial affine registration step followed by an elastic registration to handle the non-rigid deformations of the lung tissue. Both steps involved a multi-resolution strategy with 4 resolution levels for the affine procedure and 5 for the elastic. A mutual information cost function [10] is used in both cases along with a stochastic gradient descent optimizer [5]. The elastic registration deformations are modelled by a B-Spline grid [8]. The grid-size varies per resolution-level with the finest grid at the last level having a spacing of 8 voxels in each dimension.

Two different approaches to the registration procedure were evaluated: firstly registering only the anatomy within the lungs (masking out the remaining structures), and secondly registering the entire anatomy contained within the CT scans. The mask used to distinguish the lungs from other anatomy was created by means of an automatic lung segmentation procedure based on the work of Hu et al. [4].

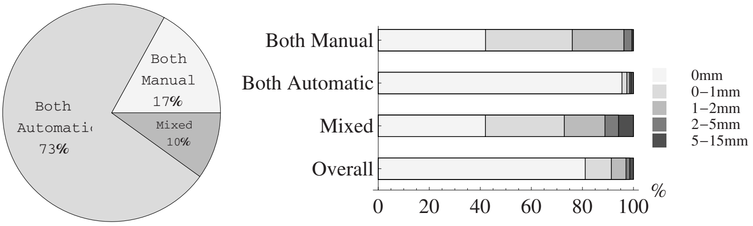
4 Results

4.1 Inter-observer Differences

The inter-observer differences were analysed to verify the ability of observers and of the system to find reproducible corresponding anatomic locations for the landmarks. The landmarks were presented to the observers in the same order, therefore in general the same points are matched manually by both observers. However, a minority of points have two different 'match-types' i.e. they are marked manually by one observer and automatically by the other. This occurred if an observer skipped a point due to difficulty with matching it manually, or where the number of manual annotations required before the system could continue automatically differed between observers. Figure 3(a) shows the proportions of all 1900 points which were matched either manually by both observers, automatically by both observers or with differing match-types (mixed). In figure 3(b) the inter-observer differences in mm are illustrated, categorised by match-type. Regardless of match-type 97% of all points had an inter-observer difference below 2mm. As is expected, points which were marked automatically by both observers are considerably more likely to have differences of 0mm than those which were marked manually, since in the automatic case a local search for the lowest SSD is performed. Points with differing match-types (mixed) tend to have larger inter-observer differences than other points. For many of these points it was difficult to find accurate matches, as evidenced by the fact that one observer considered them too difficult to match manually. For the remainder of the analysis in this work points where the interobserver difference was greater than 2mm are disregarded due to the uncertainty of the reference standard in these cases.

4.2 Registration Performance

For each image-pair the computed transform T which maps from locations in the deformed follow-up scan to locations in the original follow-up scan is applied to each of the landmark points l from the baseline scan. Since $T(l)$ may map



(a) (b)

Fig. 3. (a)The distribution over all 1900 points of various match-types. (b)Inter-observer differences categorised by match-types.

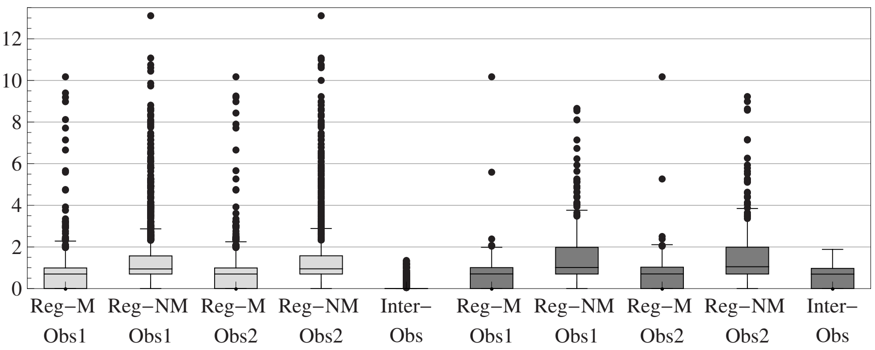


Fig. 4. Registration-observer distances δ and inter-observer differences (mm). The first five plots represent distances considering all points. The remaining five plots represent distances considering only manually matched landmarks. Reg-M and Reg-NM denote registration performed with and without lung masks respectively.

to a location between voxels, its coordinates are rounded to the nearest voxel location for comparison with the reference standard. It is clear that for an accurate registration we expect $T(l) \approx lm$, where lm is the matching point marked during reference standard formulation.

For all points lm_{obs1} marked (manually or automatically) by observer1 the distances $\delta(T(l), lm_{obs1})$ between $T(l)$ and lm_{obs1} were calculated using the appropriate T for the scan-pair. This process was performed for both T_M (the transformation from registration using lung masks) and T_{NM} (the transformation from registration with no masks). Similarly for the annotations of observer2, the distances $\delta(T_M(l), lm_{obs2})$ and $\delta(T_{NM}(l), lm_{obs2})$ were calculated.

In figure 4 box plots are presented illustrating the δ distances for each observer and for each registration procedure (with and without masks). For the registration performed using lung masks the median of the registration-observer distances is 0.7mm for both observers. Registration without masks showed an inferior result with median registration-observer distances around 0.95mm. The median inter-observer difference is 0mm due to the domination of the automatically matched points. The last five plots in figure 4 illustrate the δ and inter-observer distances taking only manually matched landmarks into consideration. Without the effect of the automatically matched points the median inter-observer distance rises to 0.69mm. It is clear that there is little difference between the distribution of inter-observer distances and that of registration-observer distances when lung masks are used during registration.

5 Conclusion

A semi-automatic system for reference standard formulation in registration has been presented. The system defines a well-distributed mesh of corresponding landmark points with limited interaction from non-expert observers. Independent observations prove the accuracy of the defined correspondence with 97% of inter-observer differences below 2mm. The efficacy of the system in quantitative analysis is proven by application to two sets of registration results, showing a significant difference between the two methods and proving the better method to have registration-observer distances in the same range as those between independent observers. Such subtle differences between registration results may easily be overlooked by evaluation techniques based on small numbers of landmark locations or on synthetically produced registration problems.

References

1. Betke, M., Hong, H., Ko, J.P.: Landmark detection in the chest and registration of lung surfaces with an application to nodule registration. Medical Image Analysis 7, 265–281 (2003)
2. Blaffert, T., Wiemker, R.: Comparison of different follow-up lung registration methods with and without segmentation. In: Proc. SPIE, vol. 5370, pp. 1701–1708 (2004)
3. Bookstein, F.L.: Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. IEEE Trans. PAMI 11, 567–585 (1989)
4. Hu, S., Hoffman, E.A., Reinhardt, J.M.: Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. IEEE Trans. Med. Imaging 20(6), 490–498 (2001)
5. Klein, S., Staring, M., Pluim, J.P.W.: Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. IEEE Trans. Med. Imaging 16(12), 2879–2890 (2007)
6. Likar, B., Pernuš, F.: Automatic extraction of corresponding points for the registration of medical images. Medical Physics 26, 1678–1686 (1999)
7. McClelland, J., Blackall, J., Tarte, S.: A continuous 4D motion model from multiple respiratory cycles for use in lung radiotherapy. Medical Physics 33 (2006)
8. 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)
9. Saff, E.B., Kuijlaars, A.B.J.: Distributing Many Points on a Sphere. The Mathematical Intelligence 19, 5–11 (1997)
10. Thèvenaz, P., Unser, M.: Optimization of mutual information for multiresolution image registration. IEEE Trans. Image Proc. 9(12), 2083–2099 (2000)
11. Urschler, M., Kluckner, S., Bischof, H.: A Framework for Comparison and Evaluation of Nonlinear Intra-Subject Image Registration Algorithms. In: IJ - 2007 MICCAI Open Science Workshop (2007)
12. Vik, T., Kabusa, S., von Berg, J., Ens, K., Dries, S., Klinder, T., Lorenz, C.: Validation and comparison of registration methods for freebreathing 4D lung CT. In: Proc. SPIE (2008)