Technical Note: A method for quality assurance of landmark sets for use in evaluation of deformable image registration accuracy of lung parenchyma

Christopher L. Guy,^{a)} Elisabeth Weiss, and Nuzhat Jan

Department of Radiation Oncology, Virginia Commonwealth University, Richmond, VA 23298, USA

Gary E. Christensen

Department of Electrical and Computer Engineering, Department of Radiation Oncology, University of Iowa, Iowa City, IA 52242, USA

Geoffrey D. Hugo

Department of Radiation Oncology, Washington University in St. Louis, St. Louis, MO 63110, USA

(Received 6 August 2018; revised 4 December 2018; accepted for publication 4 December 2018; published 28 December 2018)

Purpose: To develop a quality control method to improve the accuracy of corresponding landmark sets used for deformable image registration (DIR) evaluation in the lung parenchyma.

Methods: An iterative workflow was developed as a method for quality assurance of landmark sets. Starting with the initial landmark set for a given image pair, a landmark-based deformation was applied to one of the images. A difference image and a color overlay were generated using the deformed image and the other image of the pair. Inspection of these generated images at locations of landmarks allowed for the identification of misplaced landmarks. The observer responsible for creating the initial landmark set was tasked with review and revision of points flagged by the quality assurance procedure. Using the updated landmark sets, the process was repeated until all points were acceptable to the reviewer.

Results: Eighteen landmark sets, containing a mean (SD) of 170 (31) landmarks, were created using CT images from non-small cell lung cancer patients exhibiting large geometric changes and atelectasis resolution, making landmark specification challenging. Following the quality assurance procedure, the final landmark sets contained a mean (SD) of 165 (25) landmarks, as points too difficult to match were removed and points were added to regions deficient in landmarks. For landmark sets in which changes were made, maximum and mean differences in landmark positions before and after quality assurance ranged between 8.7–81.5 mm and 0.3–9.6 mm, respectively.

Conclusions: An effective method for improving the accuracy of landmark correspondence was presented. This quality assurance approach enables more accurate evaluation of DIR for lung parenchyma in clinical image pairs in the absence of a ground truth deformation and may be applicable to other feature-rich anatomical sites. © 2018 American Association of Physicists in Medicine [<https://doi.org/10.1002/mp.13336>]

1. INTRODUCTION

Deformable image registration (DIR) has become ubiquitous in the clinical radiation oncology setting with applications including automatic segmentation, functional imaging, dose accumulation, and adaptive radiotherapy.^{1–3} DIR capabilities, both user-guided and fully automatic, are now available in all commercial treatment planning systems and major radiotherapy software suites. The accuracy of DIR applications, dose accumulation and treatment plan adaptation in particular, depends fundamentally on the uncertainty of the underlying DIR algorithm. The lack of ground-truth transformation between a pair of clinical images makes assessment of registration accuracy a nontrivial exercise.⁴ Development and commissioning of deformable registration algorithms for clinical use require quantitative measures of accuracy on data representative of scenarios encountered during clinical practice. Additionally, when dose accumulation is involved, understanding the correlation between registration error and dose propagation uncertainty is crucial

as errors in dose could lead to detrimental clinical decisions and dose delivery.

The best measure of accuracy for DIR for clinical data is landmark registration error calculated using physician-specified points.⁵ A landmark is a pair of points, one in the fixed image and one in the moving image, which have been confirmed to physically correspond to the same anatomical location by an expert observer. The fixed landmark set can be moved to the moving image space by applying the resulting transformation of the registration. Then, the difference in position between a transformed fixed point and its corresponding moving point is defined as landmark error.

Using landmarks as an accuracy measure provides clinically meaningful validation of registration correspondence.^{6,7} However, the point matching process is time-consuming and prone to interobserver variability; placement ambiguity along curves, on smooth surfaces, and in homogeneous intensity regions; discretization errors; and, most importantly, observer error.^{8,9} Care must be taken to use landmarks distributed throughout the entire image or region of interest being

registered. This work specifically aims to address the issue of improving the accuracy of landmark placement in landmark set creation.

The correspondences of landmarks are taken as ground truth when evaluating registration error. Therefore, any errors in the landmarks sets will propagate to uncertainties in registration assessment. It is crucial that landmarks used to evaluate deformable registration be reviewed and their accuracy confirmed. The purpose of this work is to present a strategy for rapid assessment and improvement of the accuracy and quality of physician-specified landmark pairs used for image registration evaluation. Using a set of landmarks, one of the images is deformed to match the other using a landmark-based deformation. This deformation is used to assess landmark placement quality and is independent of any future DIR for which the landmarks may be used to validate. Points suspected of incorrect placement are then flagged for review by the observer. The process is repeated using the updated landmark set until no misplaced points are found. This quality assurance method is capable of detecting outliers, mislabeled points, and even small

mismatches. To demonstrate feasibility and utility of the landmark set improvement method presented here, this work focuses on landmarks of the lung parenchyma due to the feature-rich nature of the anatomical site.

2. MATERIALS AND METHODS

2.A. Patient datasets

Images were obtained for a cohort of 18 locally advanced non-small cell lung cancer patients, stages IB–IIIB treated with mean (SD) prescription dose of 63 (5) Gy, enrolled in various imaging protocols in which multiple helical four-dimensional free-breathing and three-dimensional breath hold CT images were taken during the course of radiotherapy. Longitudinal image pairs were selected, with the baseline scan acquired at the time of treatment planning and the mid-treatment scan taken after 4 (2) weeks of radiotherapy. In-plane image resolution ranged from 0.98 to 1.37 mm and slice thickness varied between 2 and 3 mm across patients. In addition to being longitudinally separated, the images used

FIG. 1. Example of landmarks for an image pair. Representative coronal slices of a single image pair are shown illustrating the typical distribution of landmarks (crosshairs) in both the baseline scan (rows 1 and 2) and the mid-treatment scan (rows 3 and 4). Baseline point locations were automatically detected by the *isi-Match* software, while mid-treatment points were matched by an expert observer. Full resolution of lobar atelectasis of the right lower lobe occurs between the baseline and mid-treatment images, inducing large geometric changes and making annotation of right lung landmarks challenging. [Color figure can be viewed at wileyonlinelibrary.com]

for this study were particularly challenging to accurately annotate as all patients presented with atelectasis at the start of radiotherapy. Furthermore, most patients exhibited large

geometric changes between baseline and mid-treatment images due to atelectasis resolution which dramatically altered appearance and location of lung parenchyma. Ground-truth correspondences between image pairs were necessary for evaluation of DIRs performed with the images reported elsewhere.¹⁰ Creation of these landmark sets is described in the following sections.

2.B. Initial landmark set generation

To facilitate the creation of landmark sets, software developed by the Imaging Sciences Institute of the University Medical Center Utrecht (Utrecht, the Netherlands) called *isiMatch* (<https://www.isi.uu.nl/research/software/isimatch/>) was used.¹¹ The software is freely available on the institution's website and packages together several key tools: an automatic point generator, a GUI for annotating images, and a semi-automatic matching feature. The *isiMatch* software's automatic point generator was used to identify highly distinct and well-distributed points throughout the lung volume of the first image of the pair which could be reliably matched to corresponding points in the second image by an observer. The initial point sets contained a mean (SD) of 170 (31) distinctive points across the 18 patients. Using the generated list of distinct points, the software's annotation GUI was used to identify the corresponding points, thus creating the landmark sets. Landmark sets were created for all 18 image pairs using the *isiMatch* software. The process of annotating the corresponding point locations for any given patient was time-intensive, ranging anywhere from 1 to 3 h, with the difficulty being highly dependent on the degree of anatomical changes present between the images. The automatic point matching feature was not available to observers for any of the 18 patients, due to the large geometric changes present in the

FIG. 2. Workflow of landmark quality assurance process. An initial landmark set is used to deform one of the images of the pair to match the other using a thin-plate spline deformation. The images are then evaluated to detect points suspected of inaccurate placement based on difference and color overlay images of the landmark-deformed image with the other image of the pair. Suspected points are then reviewed by the observer and adjusted appropriately. The process repeats until all points are found to be acceptable during the quality evaluation. The process results in a quality-assured landmark set.

FIG. 3. Complementary color overlay as a qualitative assessment of registration error. The baseline (magenta) and mid-treatment (green) images are shown before (left) and after (right) landmark-based deformation of the mid-treatment image (color in online version only). Voxels displayed in grayscale indicate similar intensities for both images while color regions are mismatched. [Color figure can be viewed at wileyonlinelibrary.com]

patient image sets; therefore, all points were matched manually. An example of landmark pairs is shown for a single patient in Fig. 1.

2.C. Quality assurance procedure

A landmark set quality assurance procedure was developed using an open-source software to identify misplaced points and to allow for the iterative improvement of landmark set quality. An overview of the workflow is illustrated in

Fig. 2. As the first step, the mid-treatment image of the image pair is deformed via transformix using a thin-plate spline (TPS) deformation constructed from the initial landmark set. The deformed mid-treatment image which results from this deformation is expected to be well-aligned to the pretreatment image at locations in close proximity to landmarks, provided that the given point pair is accurately matched. An example of an image set deformed in this manner is illustrated in Fig. 3. Using both a color overlay and a difference image between the deformed mid-treatment and baseline images,

FIG. 4. Example of landmark quality assurance. Landmark point locations are indicated by crosshairs. The vessel on which point B is located is clearly misaligned as indicated in both the difference image (left) and the complementary color overlay (right). As a result, point B would be flagged for closer review by the physician performing the annotation while point A would be considered adequately matched. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE I. Results of landmark quality assurance procedure.

Patient	Number of landmarks					Difference in landmark position (mm)		
	Initial	Final	Removed	Changed	Added	Max	Mean	SD
1	193	179	16	93	2	43.2	4.5	8.6
2	146	146	0	0	0	—	—	—
3	198	187	15	32	4	26.4	1.3	4.4
4	207	188	29	94	10	81.5	9.6	18.1
5	177	176	1	28	0	17.3	0.9	2.7
6	160	177	2	99	19	42.4	2.3	4.8
7	161	143	21	17	3	22.6	0.6	2.5
8	178	166	19	20	7	8.7	0.3	1.0
9	149	149	0	0	0	—	—	—
10	181	180	2	100	1	19.5	2.3	3.4
11	106	117	8	93	19	23.6	3.9	4.4
12	180	175	9	15	4	21.1	0.7	2.8
13	193	197	0	12	4	15.6	0.4	2.0
14	124	124	0	0	0	—	—	—
15	200	195	10	41	5	14.5	0.8	2.0
16	225	194	39	45	8	21.1	1.0	2.7
17	131	131	0	0	0	—	—	—
18	150	150	0	0	0	—	—	—

points suspected of erroneous placement are flagged for review by the original observer. Misplaced points are indicated by color voxels, as opposed to grayscale, in the color overlay and significant intensity patterns in the difference image within the immediate vicinity of the given landmark. A demonstration case of the review process is shown in Fig. 4. Visible misalignment of high-intensity structures in the color overlay image and non-zero intensities in the difference image in close proximity to a landmark point would cause that point to be flagged. Once the placement of all flagged points is confirmed to be correct or adjusted by the observer, the updated landmark set is used to repeat the process. The quality assurance procedure continues until all points are deemed accurate.

This quality assurance method was developed using an open-source software. A software package for image registration, elastix (<http://elastix.isi.uu.nl/>), also comes with a tool called transformix which enables the user to analyze and apply registration transformations to images and point sets.¹² Code to convert the output point pairs file of the isiMatch software, used to generate the initial landmark sets, to a landmark-based TPS transformation input parameter file for transformix was developed and has been made freely available online (<https://github.com/guycl/landmarkQA>). The TPS deformation model combines an affine deformation with

local deformation and is calculated based on a given set of corresponding points, landmarks in this case. Corresponding points are matched exactly using a TPS model. Finally, the open-source visualization software 3DSlicer (slicer.org) was used to display both the difference image and overlay images. The landmark converter code also comes with an option to output landmarks as 3DSlicer-compatible fiducial files, enabling visualization of the landmarks with their images.

3. RESULTS

The revised point sets contained a mean (SD) of 165 (25) distinctive points across the 18 patients. A number of different annotation errors were found using the quality assurance procedure outlined above. For the 18 landmark sets, the results of the landmark quality assurance process are shown in Table I. Five landmark sets were accepted as accurate without modification. These landmark sets corresponded to patient images showing little or no change in atelectasis volume and were therefore less challenging to annotate compared to the other image sets of this study. The remainder of sets underwent multiple rounds of quality assurance which concluded with all point pairs being either accepted or removed if deemed too challenging to match. Of the landmark sets which needed some degree of revision, an average of 53, 13, and 7 point pairs

FIG. 5. Landmarking error: incorrect vessel. An example of selection of the wrong vessel tree is shown. (a) The baseline point (far left), the initially chosen corresponding point in the mid-treatment image (top row), and the revised corresponding point (bottom row). The difference and complementary color overlay of the image pair after landmark-based deformation were used to detect the error in placement based on the region around the baseline landmark position (circles). (b) The initial point was located three axial slices away from the corrected point, corresponding to a displacement of more than 6 mm. Arrows highlight the trajectory of the initially chosen vessel back to the slice of the revised point. Landmark point locations are indicated by crosshairs. [Color figure can be viewed at wileyonlinelibrary.com]

were changed, removed, and added, respectively. Maximum and mean changes in landmark position ranged between 8.7–81.5 mm and 0.3–9.6 mm, respectively, for the same subset.

The most severe errors were the result of the observer selecting the wrong vessel tree, as shown in Fig. 5. Because of the longitudinal nature of this dataset, coupled with the large deformations often present due to changes in atelectasis, spatial overlap of different vessel branches was common. In this example, two vessel trees similar in appearance and near the position of the initial point crossed each other in the axial plane. The observer selected the incorrect vessel tree of the two on which to place the corresponding point. The quality assurance procedure easily identified this mismatch in alignment, allowing the observer to revise the chosen point and improve accuracy of the landmark set. The most common error identified was the selection of the wrong bifurcation along the same vessel tree. An example is shown in Fig. 6, where the observer-chosen point was moved further down the vessel tree to obtain an accurate match. Smaller errors were also caught such as placement of a corresponding point on the incorrect side of a large vessel, as shown in Fig. 7. In this example, the correct vessel structure is matched by the observer, but the incorrect edge was chosen. While only a few millimeters difference in location, selection of the wrong vessel

side generates obvious misalignment of the difference image and color ghosting (i.e., incomplete overlap of corresponding magenta and green structures) in the overlay image. The wide range of errors detectable by this quality assurance method enabled significant improvement in the accuracy of landmark sets.

The quality assurance procedure also allowed for identification of regions where additional landmarks were needed. Locations far from any landmarks were not expected to show perfect alignment between the landmark-deformed mid-treatment image and the baseline scan, especially if large deformations occurred to the underlying anatomy. To assure that registration landmark error calculated based on the landmark sets was representative of registration error throughout the entire region of interest, points were manually added to regions lacking landmarks and showing misalignment during review. The additional points were added to the landmark set and were matched by the same physician–observer, providing further quality improvement.

4. DISCUSSION

In this work, a method for rapid assessment and improvement of landmark set quality is presented. While the use of

FIG. 6. Landmarking error: incorrect bifurcation. An example of selection of the wrong bifurcation is shown. (a) The baseline point (far left), the initially chosen corresponding point in the mid-treatment image (top row), and the corrected corresponding point (bottom row). The difference and complementary color overlay of the image pair after landmark-based deformation were used to detect the error in placement based on the region around the baseline landmark position (circles). (b) The initial point was located three axial slices away from the corrected point, corresponding to a displacement of more than 6 mm. Arrows highlight the trajectory of the initially chosen vessel back to the slice of the revised point. Landmark point locations are indicated by crosshairs. [Color figure can be viewed at wileyonlinelibrary.com]

FIG. 7. Landmarking error: wrong side of vessel. The baseline point (far left), the initially chosen corresponding point in the mid-treatment image (top row), and the corrected corresponding point (bottom row) are shown as an example of erroneous landmark placement detectable by the landmark quality assurance method. The wrong side of the vessel was chosen as the corresponding point, causing noticeable discrepancies in the quality assurance images despite the error in placement being minimal in magnitude (~2 mm). The difference and complementary color overlay of the image pair after landmark-based deformation were used to detect the error in placement based on the region around the baseline landmark position (circles). Landmark point locations are indicated by crosshairs. [Color figure can be viewed at wileyonlinelibrary.com]

landmark-based deformations has been utilized previously to predict landmark locations and automatically perform the matching, for example, *isiMatch*, our method is novel in that it uses a landmark-based transformation to deform intensity images and assess integrity of user-specified locations. Additionally, tools to perform the outlined method are either open-source or have been made freely available, allowing readers to easily implement the method in their own clinical practice or research setting.

Well-known limitations of landmark sets for image registration quality assurance include the inability to accurately specify points in regions of homogeneous intensity or on feature-deprived surfaces. Sliding interfaces also present a major challenge to image registration in general. Overcoming these limitations was not the goal of this work; instead, we focus on landmark sets created in feature-rich volumes, namely healthy lung parenchyma.

Even within the lungs, extensive time and effort are required to specify a sufficient number of matching voxels in feature-rich locations. Automatic landmark matching methods such as the scale invariant feature transformation can potentially reduce the effort necessary to specify correspondences,^{13,14} though point detection in the presence of large deformations may remain a challenge as demonstrated by *isiMatch*'s performance in the creation of landmarks for this work. Our approach can easily be used for quality assurance of automatic landmarking approaches. Additionally, this method is applicable to any image modality granted that both images of the pair are of the same modality, as the landmark-based deformation does not rely on image intensity. Utilizing images with adequate contrast, such as MRI instead of CT for soft-tissue regions of interest, would extend the applicability of this method.

The importance of quality assurance of landmark sets increased with the difficulty of the registration between the given image pair. The patients used for this study exhibited large geometric changes and complicated deformations, explaining the large differences in point positions between the initial and final landmark sets. Incorrect overlap of vessel trees in the lung, as observed for multiple patients of this study, also necessitated careful review of landmark sets prior to use. The quality assurance method presented here provides an efficient manual review of results to identify incorrectly placed points. Future work could automate the quality assurance process to determine points suspected of misplacement based on features of the difference image which are visually quite pronounced and potentially detectable by an algorithm. For example, a similarity metric such as the sum of squared intensities of the difference image could be calculated in a localized region about each landmark point to quantify the degree of mismatch. A threshold on this metric could then be used to automatically flag points with suspected misplacements.

5. CONCLUSION

Accuracy of landmark sets was evaluated and improved through the development of a novel quality assurance procedure. Through the iterative process, points with questionable accuracy were identified and reviewed by the observer. Reduced uncertainty and observer error in placement of landmarks allows for the accurate evaluation of DIR error for clinical image pairs in the absence of a ground truth deformation.

ACKNOWLEDGMENTS

This work was supported by a research grant from the National Cancer Institute of the National Institutes of Health under award number R01CA166119. The authors thank Leonid B. Reshko and Elizabeth Jaworski for their contributions to landmark set creation.

CONFLICTS OF INTEREST

Elisabeth Weiss receives research support from Varian Medical Systems and the National Institutes of Health and receives royalties from UpToDate. Gary Christensen maintains research grants from National Institutes of Health and has received a gift from Roger Koch to support research. Geoffrey Hugo receives research support from the National Institutes of Health and Varian Medical Systems.

^{a)}Author to whom correspondence should be addressed. Electronic mail: christopher.guy@vcuhealth.org; Telephone: +(804) 628-2423; Fax: +(804) 827-1861.

REFERENCES

1. Czajkowski P, Piotrowski T. Registration methods in radiotherapy. *Rep Pract Oncol Radiother*. 2019;24:28–34.
2. Oh S, Kim S. Deformable image registration in radiation therapy. *Radiat Oncol J*. 2017;35:101–111.
3. Simon A, Nassef M, Rigaud B, et al. Roles of deformable image registration in adaptive RT: from contour propagation to dose monitoring. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Milan: IEEE; 2015:5215–5218.
4. Rivest DRC, Riauka TA, Murtha AD, Fallone BG. Assessment of a commercially available automatic deformable registration system. *J Appl Clin Med Phys*. 2010;11:3452.
5. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: report of the AAPM Radiation Therapy Committee Task Group No. 132. *Med Phys*. 2017;44:e43–e76.
6. Kanai T, Kadoya N, Ito K, et al. Evaluation of accuracy of B-spline transformation-based deformable image registration with different parameter settings for thoracic images. *J Radiat Res*. 2014;55:1163–1170.
7. Brock KK. Results of a multi-institution deformable registration accuracy study (MIDRAS). *Int J Radiat Oncol*. 2010;76:583–596.
8. Varadhan R, Karangelis G, Krishnan K, Hui S. A framework for deformable image registration validation in radiotherapy clinical applications. *J Appl Clin Med Phys*. 2013;14:4066.
9. Kashani R, Hub M, Kessler ML, Balter JM. Technical note: a physical phantom for assessment of accuracy of deformable alignment algorithms: phantom for assessment of deformable alignment. *Med Phys*. 2007;34:2785–2788.
10. Guy CL, Weiss E, Christensen GE, Jan N, Hugo GD. CALIPER: a deformable image registration algorithm for large geometric changes during radiotherapy for locally advanced non-small cell lung cancer. *Med Phys*. 2018;45:2498–2508.
11. Murphy K, van Ginneken B, Klein S, et al. Semi-automatic construction of reference standards for evaluation of image registration. *Med Image Anal*. 2011;15:71–84.
12. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: a toolbox for intensity based medical image registration. *IEEE Trans Med Imaging*. 2010;29:196–205.
13. Vickress J, Battista J, Barnett R, Morgan J, Yartsev S. Automatic landmark generation for deformable image registration evaluation for 4D CT images of lung. *Phys Med Biol*. 2016;61:7236–7245.
14. Paganelli C, Peroni M, Baroni G, Riboldi M. Quantification of organ motion based on an adaptive image-based scale invariant feature method: quantification of organ motion based on adaptive-SIFT. *Med Phys*. 2013;40:111701.