Accurate Landmark Pairs Detection for 4DCT Lung Deformable Image Registration Verification

Yabo Fu, Xue Wu, Harold H. Li, Deshan Yang*

Department of Radiation Oncology, Washington University in Saint Louis

*Corresponding author:

Deshan Yang

Washington University in Saint Louis

yangdeshan@wustl.edu,

Abstract

Purpose: To automatically and precisely detect large quantity of landmark pairs between pairs of intra-patient image volumes for the purpose of deformable image registration (DIR) verification. We expect that the generated landmark pairs will significantly augment the current DIRLAB benchmark datasets in both quantity and positional accuracy.

Methods: Large number of landmark pairs were detected within the lung between the end-exhalation (EE) and end-inhalation (EI) phases of the 10 DIRLAB 4DCT lung datasets. Foerstner operator was applied to find thousands of landmarks within the lung of the EI phase. A parametric image registration method (pTVreg) was used to register the EE and EI phases to establish the corresponding landmarks in EE. Because image registration was subject to registration errors, a Multi-Stream Pseudo-Siamese (MSPS) network was developed to further improve the landmark positional accuracy in EE by directly predicting a vector of xyz shift to optimally align the landmarks in EE to that in EI. Positional accuracy of the detected landmark pairs was evaluated using both the digital phantoms with known ground truth and the publically-available DIRLAB landmark pairs.

Results: Compared to DIRLAB datasets, the quantity of detected landmark pairs increased more than threefold. The mean and standard deviation of target registration error (TRE) was 0.47±0.45 mm with 90% of landmark pairs having a TRE smaller than 1mm for the ten digital phantom cases with known ground truths. The mean and standard deviation of TRE was 0.73±0.53 mm with 83% of landmark pairs having a TRE smaller than 1mm for the ten DIRLAB cases with 300 manual landmark pairs as the ground truths.

Conclusion: A novel method was developed to automatically and precisely detect large quantity of landmark pairs between intra-patient volumetric medical image pairs for quantitative evaluation of DIR algorithms.

Keywords: Deformable image registration, image feature detection, medical image processing, radiation therapy

1 Introduction

Deformable image registration[1,2] (DIR) is a key enabling technology for many important advanced radiotherapy techniques (e.g. adaptive radiotherapy[3-5]) and critical clinical tasks (e.g. target definition[6,7], automatic segmentation[8-10], motion estimation[11-13], dose accumulation[14-16] and treatment response evaluation[5,17,18]). DIR accuracy, which is the correspondence of matching points between two images under DIR, is often inadequate and largely dependent on the operator, DIR algorithm, implementation and image quality. It is critical to evaluate DIR algorithms quantitatively using the benchmark datasets before the accuracies could be assessed and the DIR parameters could be understood and optimized by the operators. Hyperparameters of DIR algorithms are often optimized case by case which make it necessary to develop a patient-specific validation method for different cases. The lack of gold standard to evaluate DIR algorithms has slowed the use of DIR in clinical practices [x]. Many methods have been proposed to validate the accuracy of various DIR algorithms [xxx]. These methods can be divided into two categories: 1) using real patient images with manually selected landmark pairs, and 2) using phantom images with artificial deformation vector field (DVF). Since it is impractical to build a phantom that is sophisticated enough to simulate the complicated anatomical structures and motion fields, the only trustable way to assess the accuracy of a specific DIR algorithm is to compute Target-Registration-Error (TRE) on manually selected landmark pairs of real patient images. However, the manual landmark selection process is very labor intensive if a relatively large number of landmarks are to be selected. Therefore, we are motivated to develop a fully automatic method that is capable of detecting a dense set of landmark pairs for accurate and complete validation of DIR algorithms.

A recent study shows that the DIR performance solely assessed using sparse contrast-rich features may not reflect the true DIR performance in uniformly low contrast anatomy [x]. Therefore, a dense set of landmark pairs is needed to gain a more complete understanding of registration accuracy. Castillo et al. provided a widely used 4DCT lung benchmark datasets, the DIRLAB, with 300 manually selected landmark pairs for the end-inhalation (EI) and end-exhalation (EE) phase. The limitations of the DIRLAB datasets

are the relatively small number of landmark pairs and large observers' variability. Yang et al. recently developed a method, MRICGM (Multiple-Resolution Inverse-Consistency Guided Matching), to automatically and accurately detect landmark feature pairs between intra-patient 3D images[19]. One limitation of the MRICGM is that it took the best available matching landmark in the second image as the corresponding landmark for a specific landmark in the first image. However, the location of the best available matching landmark in the second image may not be adequately close to the landmark in the first image. Murphy et al. proposed a semi-automatic method to construct DIR benchmark datasets of 47 pairs of temporal thoracic CT scans with 100 landmark pairs per case. The landmark pair correspondence was established using block-matching and manual confirmation. However, the 100 landmark pairs per case were inadequate for a complete DIR evaluation. Authors did not report the accuracy of the automatic feature matching prior to manual confirmation. Werner et al. and Polzin et al. both used a Foerstner3D operator to detect landmarks within one image and then transferred the landmarks to the other image using a cross correlation based block matching strategy. The limitations of this method are 1) the cross correlation metric is an inaccurate similarity measure between images that were subject to significant irregular deformations, 2) only a small fraction of the detected landmark candidates can be reliably transferred by this approach, 3) the accuracy of the block matching strategy was not reported.

Landmark pairs could be detected using DIR algorithms by propagating the detected landmarks between the images to be registered. However, DIR algorithms are subject to registration errors due to multiple factors including inaccurate similarity metrics, regularization methods, and inappropriate parameter selection and so on. The commonly used image similarity metrics in DIR algorithms such as sum of squared intensity difference (SSD), normalized correlation coefficient (NCC), normalized gradient field (NGF) or mutual information (MI) were not accurate enough when the images to be registered were subject to significant irregular deformations. In the case of lung registration, very few local features were stable across the EE and EI phases especially for landmarks near the diaphragm where respiratory motion is significant. Out of the 26 DIR algorithms that were reported by the DIRLAB website, the best TREs for the

ten DIRLAB datasets were on average 0.91±1.07 mm which is insufficient for the purpose of DIR accuracy validation.

In this paper, we present a new method to automatically detect large quantity of landmark pairs with excellent positional accuracy between EE and EI phases of the 4DCT lung datasets. Compared to previously published studies, the major contributions of our work are:

1) A robust landmark detection method was proposed by using the Harris-Stephens corner detection algorithm on lung vasculature tree probability maps.
2) A Multi-Stream Pseudo-Siamese (MSPS) network was designed to directly predict a vector of xyz shift to optimally align the landmarks in EE to that in EI.
3) A new outlier rejection method was used to keep only the results with robust MSPS predictions.
4) New sets of dense landmark pairs for the 10 DIRLAB datasets and 9 EMPIRE10 datasets were shared at xxx for DIR evaluations.

2 Material and methods

2.1 Materials

Ten DIR-LAB 4DCT lung cases were used in this study. For the accuracy evaluation of the proposed method, a digital phantom dataset was generated by deforming the end exhalation (EE) phase image by a generated deformation vector field (DVF) into a simulated end inhalation (EE) phase image. In this way, the voxel correspondence ground truth between the EE phase image and the simulated EI phase image are known. For each EE phase image and simulated EI phase image pair, we applied the MRICGM method to detect the matching feature pairs [19]. The positional accuracy of these landmark pairs were refined using the proposed deep learning network. TRE values calculated using these landmark pairs before and after the network refinement were compared and listed in Table 1.

2.2 Procedures to detect landmark pairs

2.2.1 Automatic landmark detection

The initial step of landmark pair detection is to automatically determine a large number of feature points in one image. Scale Invariant Feature Transform (SIFT) is one of the most commonly used method to detect landmarks [xxx]. Vickress et al. has tested three landmark detection methods including manual, SIFT and SIFT with manual editing. However, SIFT is apt to detect blob features which are generally located in smooth areas with few representative features [x]. Important diagnostic or respiratory motion information are generally reflected by vessel crossing points, vascular endpoints and tissue boundary points. To detect more feature points, Yang et al. employed both a 3D SIFT feature detector and a 3D Harris-Laplacian corner detector. Similarly, Zhang et al. proposed a hybrid method which was based on Harris and SIFT to effectively detect lung features. Foerstner3D operator was also used to detect landmarks for DIR validation [xx]. Murphy et al. selected the landmarks based on a distinctiveness value which was calculated using the image gradient. However, these methods were applied directly to the original CT images which were subject to motion artifacts and noise. As a result, the detected landmarks may have positional uncertainties which could lead to poor measurement precision and accuracy. The positional uncertainties of landmarks could lead to inaccurate TRE calculation for DIR evaluation. To overcome this limitation, we propose to detect feature points [x] on the pulmonary vasculature probability maps that were generated from the original CT images. The probability maps were produced using a stacked multiscale feature learning model [x]. This model could make probabilistic predictions on the occurrence of pulmonary vessels by feeding the learnt voxel-wise features to a logistic regression classifier. The pulmonary vasculature structures could provide abundant information about the vessels and reflect the motion of the lung accurately. The vasculature probability maps were used as robust surrogates to the original CT images. Harris-Stephens algorithm [x] was adopted to detect feature points on the vasculature probability maps. In the post-processing step, landmarks that were either with poor contrast or within 3 mm distance to its surrounding landmarks were eliminated to make sure that the remaining landmarks were visually identifiable and well distributed throughout the

lung. Our preliminary results had shown that this method could robustly detect large number of feature points with high quality. The procedures of the landmark detection process were shown in Fig. x. Fig.2 shows one example of the segmented vessels and the vasculature probability map with detected landmarks.

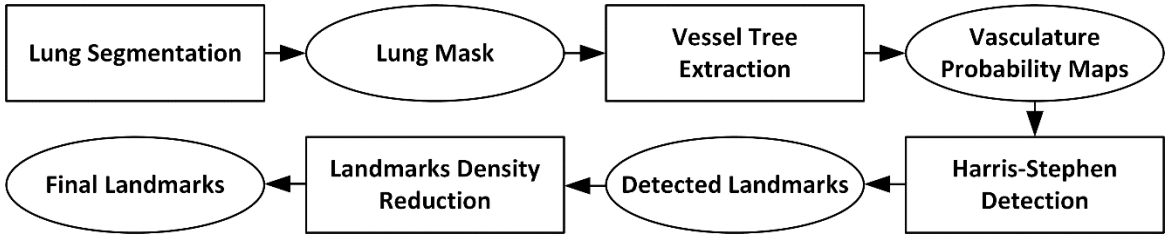

Fig 1. Procedures to detect lung feature points

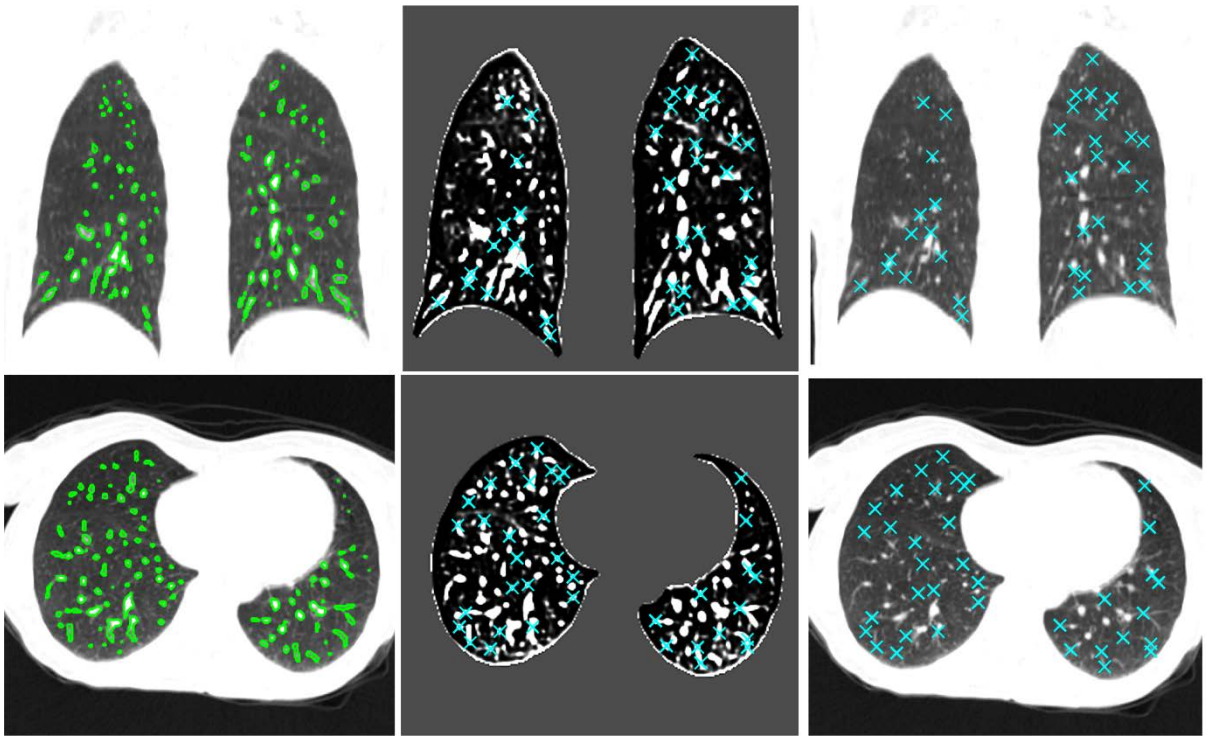

Fig 2. Detected landmarks were shown by cross marks. Left: Original CT image with vessels, Middle: Vasculature probability maps with landmarks, Right: Original CT image with landmarks.

2.2.2 Establishing landmark correspondence

Multiple methods have been proposed to establish landmark correspondence [xx]. Yang et al. detected tens of thousands of feature points separately on two images. These landmarks were subsequently paired

using a multi-resolution guided matching scheme. The matching process was based on a descriptor matching confidence value which was calculated as the dot product of two corresponding feature descriptors. Murphy et al. proposed a semi-automatic landmark matching process. An initial estimate of the corresponding landmark location was manually annotated by one observer. A local block-matching search method was used to improve upon this initial estimate. The point with minimal SSD within the region of interest was selected as the final estimate after the block-matching search. Werner et al. used a cross correlation based block matching strategy to automatically establish landmark correspondence between images. The MRICGM proposed by Yang et al. was reported to have TRE of 0.72±0.67mm for three 4DCT lung phantoms. No accuracy evaluation was reported for the above mentioned block-matching methods.

In authors' observation, the lack of accurate similarity measures between image patches is one of the main causes that has limited the accuracy of image patch matching methods. Conventional similarity metrics such as SSD, NCC and MI were not able to accurately reflect the distances between image patches that were subject to significant deformations. To improve the accuracy of similarity metrics, various metrics have been proposed using both machine learning [xx] and deep learning methods [xx]. Supervised machine learning methods have been successfully applied to learn patch descriptors [xxxx]. These methods were able to outperform hand-crafted approached [x]. Simonovsky et al. proposed a deep metric for multimodal image registration. They trained the network from scratch using a few aligned image pairs. They demonstrated that the trained deep metric was able to outperform MI by a significant margin. Simo-Serra et al. developed a CNN to learn discriminant patch representations which were proved to generalize well against scaling, rotations and non-rigid deformations. The discriminant patch representation was a 128-D descriptors whose Euclidean distances reflect patch similarity. Haskins et al. developed a deep learning based similarity metric for 3D MR-TRUS registration. Similarly, Sedghi et al. trained a semi-supervised deep metrics for MRI T1-T2 image registration. Zagoruyko et al. explored and compared a variety of different CNN models for image patches comparison tasks. Triplet networks have been explored for image feature learning and retrieval [xxxx].

In this study, we choose to use a two-step process to propagate the landmarks between images. Firstly, the initial landmark correspondence was established using a parametric DIR algorithm. Secondly, the positional accuracies of the landmark pairs were refined using a Multi-Stream Pseudo-Siamese (MSPS) network. The MSPS network implicitly learnt an optimal similarity metrics and predicted of a vector of xyz shifts which was used to accurately align the image patches.

1) Initial landmark correspondence

To establish initial landmark correspondence between images, we chose to use the PTVreg [x] algorithm which is one of the best performing DIR algorithms reported by the DIRLAB website. The PTVreg was a parametric image registration method with total variation regularization. Advantages of the PTVreg include 1) it supports non-smooth displacement that occurs at the pleural cavity; 2) its performance is robust to parameter selection; 3) it could provide accurate registration results for the whole lung. The PTVreg algorithm was adequate to closely locate the corresponding landmarks in the second image. Assessed using the DIRLAB 300 manual landmark pairs, the PTVreg has an average TRE of 0.92±1.06 mm for the ten 4DCT lung cases. DIR algorithms were subject to registration errors which were caused by multiple factors including inaccurate similarity metrics, regularization methods, and inappropriate parameter selection and so on. The positional accuracies of the landmark pairs at this step were inadequate to evaluate DIR algorithms.

2) Landmark positional accuracy refinement using MSPS

CNN methods have been very successfully in various classification tasks such as image identification, image segmentation [xxx]. Most networks were designed to tolerate spatial variations through the use of spatial pooling layers and data augmentation techniques. Recently, Jaderberg et al. presented a Spatial Transformer Networks (STN) for resolving, instead of tolerating, image misalignment. They integrated image warping operation within STN and showed that such operations were sub-differentiable. They applied STN on affine image registration using the MNIST digit datasets. Multiple CNNs were proposed subsequently for rigid [x] and non-rigid [xx] image registrations using similar concepts. In this study,

we reformulate the problem of image patch matching as a rigid registration problem with only 3 degree of freedoms which are translations in x, y, z directions.

a) Network configuration

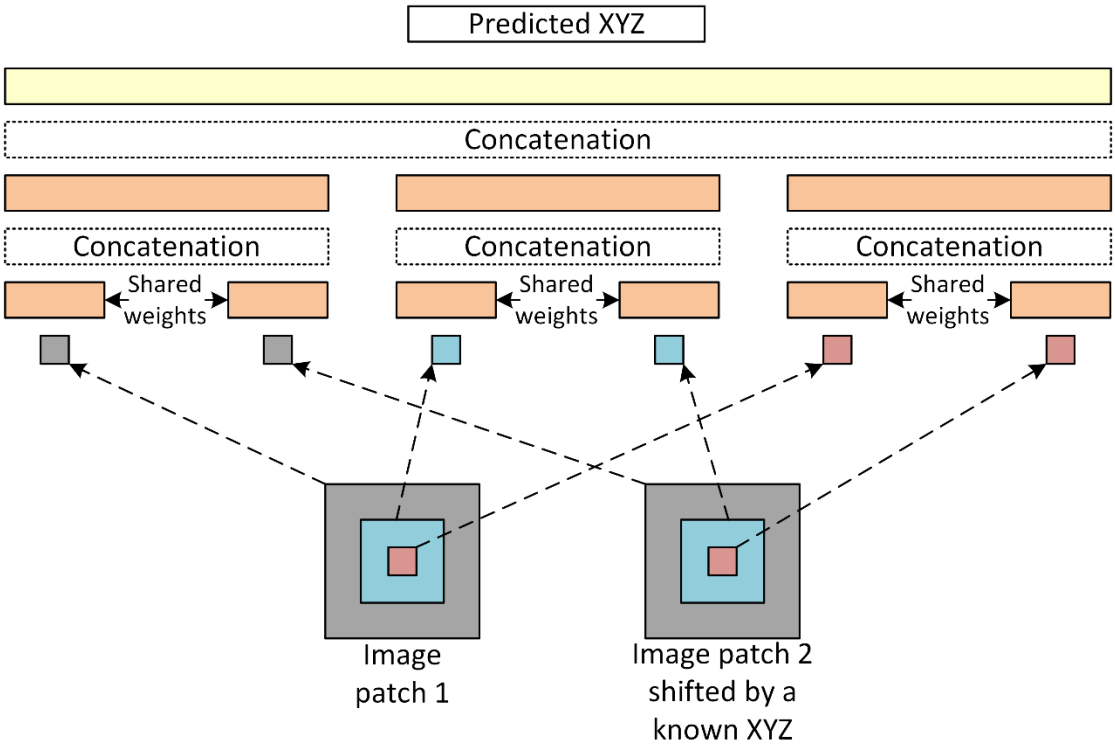

Fig. 3 The MSPS network architectures. A central-surround three stream network uses a siamese type architecture for each stream. Color code: orange = Conv+Relu, yellow=fully connected layer.

The architecture of the proposed MSPS network is shown in Fig. 3. It consists of three separate streams with each stream focusing on images with different patch sizes. The image patch sizes for the three streams were 32×32×32, 16×16×16 and 8×8×8. These image patches were all centered on the landmark pair to be refined. Each stream has its own dedicated layers before joining the fully connected layer. In each stream, the two image patches were first processed by 7 convolutional and Relu (Conv+Relu) layers with shared weights. These shared layers encoded the image patches into a common space where they were further compared. The encoded image patches were concatenated and processed by another 7 Conv+Relu layers to generate the deep similarity descriptor of the two image patches. The last concatenation operation

requires the shapes of deep similarity descriptors of the three streams to be consistent. Different stride values of 4, 2, and 1 were used for the large, medium and small image patches respectively at the first convolutional layers of each stream to generate similarity descriptors with consistent shapes. The deep similarity descriptors of the three streams were finally concatenated to form the 'multiscale deep similarity descriptors' which were fed to the fully connected layer for final prediction. The 'multiscale deep similarity descriptors' was a vector with a length of ~15,000. The network has a total of ~7 million parameters whereas ~4 million of parameters belongs to the fully connected layers.

The choice of the three streams with different patch sizes was because 1) the stream with large image patch size help increase the network robustness by using a large reception field; 2) the streams with medium and small image patch sizes help the network to focus more on the fine details near the center; 3) the central part of a patch was considered three times which implicitly puts more emphasizes on the pixels closer to the center of a patch and helps to improve the precision of the matching. Batch normalization was not used because the artificially generated training image patches have a different covariance shift from the PTVreg generated image patches. Our experimental results with batch normalization have demonstrated that the use of batch normalization will reduce the accuracy of the network predication when applied to PTVreg generated image patches. Pooling layers which are commonly used to tolerate spatial variance were not used since the network aims to do exactly the opposite which is to resolve the spatial variance. No dropout layers were used in the network since the network is designed to perform deterministically when predicting the xyz values.

b) Datasets preparation

The ten DIRLAB datasets have a total of 3000 manual landmark pairs with 300 landmark pairs for the EI and EE phases per case. Landmark pairs of cases 1-8 were used as training datasets. Landmark pairs of case 9 were used as validation datasets. Landmark pairs of case 10 were used as testing datasets. The images at EI and EE phases were resampled to isotropic resolution in the preprocessing step. The MSPS network aimed at refining the positional accuracies of landmark pairs that were generated using the PTVreg

registration results. PTVreg has an average TRE of 0.92±1.06 mm for the ten 4DCT lung cases, which suggests that the TREs of the majority of the initial landmark pairs were under 4mm. To simulate the TRE distribution of the PTVreg, we prepared the training datasets by artificially shifting one of the well-aligned image patches by known xyz values. The x, y, and z values were separately sampled from a normal distribution with zero mean and 2mm standard deviation. The training images were arranged in a 4D array of size with the fourth dimension being the image pair. For each image pair, one additional image pair was created by reversing the order of the fourth dimension and changing the sign of the known xyz values. This data augmentation operation was specifically designed to train the network to be anti-commutative:

$$f(a,b) = \overrightarrow{xyz} = -f(b,a)$$

where f represents the network forward prediction operation, a and b denote the two image patch to be matched respectively, $\overrightarrow{xyz}$ denote the predicted xyz vector.

c) Training configurations

The network was implemented using Tensorflow [x] in Python. The training parameters were initialized randomly using a Gaussian distribution (mean = 0, std = 0.01). The network was trained using Adam optimizer with a constant learning rate of 1e-4. The loss function of the network was chosen as the mean squared error between the predicted xyz values and the ground truth xyz values. The learning process was assessed using the 300 validation manual landmark pairs of case 9 of the DIRLab datasets at every 500 iterations. The training process was stopped if the loss did not improve for 5 consecutive validations, i.e. 2500 iterations. A GeForce GTX 1080 Ti GPU with 11GB RAM and 3584 CUDA cores was used for training.

3) Outlier rejection

An outlier rejection method was introduced to reject inaccurate network predictions. The assumption is that the network predictions should be consistent if the centers of the two image patches were both shifted by a same small vector:

$$f\left(a_{p_1}, b_{p_2}\right) = f\left(a_{p_1+w}, b_{p_2+w}\right) \; with \; w = -1,0,1$$

where f represents the network forward prediction operation, a_{p_1}, b_{p_2} represent image patches centered on one landmark pair p_1, p_2 respectively, w is the small vector which is used to shift the image patches. For a certain landmark pair, the network performed 27 predictions within a 3×3×3 neighborhood. Another 27 predictions were calculated by reversing the order of the image patches. Therefore, a vector of 54 predictions were generated for each landmark pair. If the network is able to perform robustly and accurately at the landmark pair, the 54 predictions should be consistent with one another. However, the trained network was not able to perform robustly on every landmark pair that we were trying to refine. One of the reasons was that only 2700 landmark pairs were available to train the network, which makes it difficult for the network to generalize well on every possible landmark pair and image patches. The inconsistency of the 54 predictions were described by a vector:

$$E = \left\| f\left(a_{p_1}, b_{p_2}\right) - f\left(a_{p_1+w}, b_{p_2+w}\right) \right\| \; with \; w = -1,0,1$$

The mean and variance of E, $\mu(E) and \; \sigma^2(E)$, were calculated and used to reject inaccurate network predictions. In this study, we chose to use the 75th percentile values of $\mu(E) and \; \sigma^2(E)$ as the thresholding values. An alternative is to use constant values as the thresholding values. For the datasets that we have experimented on, 75th percentile were ~0.5mm for $\mu(E)$ and ~0.1mm for $\sigma^2(E)$. After the outlier rejection, the final prediction for a certain landmark pairs was taken to be the mean value of the 54 predictions.

3 Results

3.1 Accuracy evaluations

Ten digital phantoms were generated by deforming the EE phase images using known DVFs to evaluate the accuracy of the proposed method. The known DVFs were generated using traditional Horn-Schunck optical flow method. Landmark pairs were detected between the simulated EI and real EE phases using the proposed method. TREs were reported in Table 1 for both the initial landmark pairs and the final landmark pairs. On average, 1427 landmark pairs were detected for the 10 digital phantoms. The large maximum TRE for the initial landmarks suggested the performance of PTVreg was not consistent

throughout the lung. The mean, standard deviation and maximum TREs were reduced after the MSPS network refinement. The mean and standard deviation of the TREs for the final landmark pairs were 0.47±0.45mm. The TREs were smaller than 1mm for around 90% of the final landmark pairs.

Table 1: TREs (mean±std[max]) of the proposed method on 10 digital phantoms

Case #	**Voxel size (mm^3)**	**# of landmark pairs**	**Ground truth DVF (mm)**	**TRE for landmark pairs (mm)**		**% of final landmark pairs with TRE < 1mm**
				Initial	**Final**	
1	0.97×0.97×0.97	1126	3.43±2.86[12.38]	0.46±0.39[3.32]	**0.34±0.29[1.67]**	97
2	1.16×1.16×1.16	1546	4.67±4.23[18.93]	0.44±0.32**[1.76]**	**0.35±0.37**[3.81]	95
3	1.15×1.15×1.15	1201	5.55±4.08[16.75]	**0.51±0.38[2.47]**	0.51±0.55[4.76]	87
4	1.13×1.13×1.13	874	7.55±5.11[19.94]	**0.49±0.36[2.76]**	0.55±0.54[4.86]	86
5	1.10×1.10×1.10	1077	4.91±4.84[21.49]	**0.52±0.53**[9.07]	0.58±0.60[**4.15**]	83
6	0.97×0.97×0.97	1410	9.30±7.46[38.15]	0.68±1.30[31.75]	**0.57±0.55[4.39]**	84
7	0.97×0.97×0.97	1792	8.18±6.73[29.51]	0.58±1.45[28.26]	**0.45±0.41[4.27]**	92
8	0.97×0.97×0.97	2798	8.58±6.71[30.02]	0.78±2.53[41.51]	**0.51±0.48[4.85]**	87
9	0.97×0.97×0.97	855	5.81±3.77[14.49]	0.45±0.37[**2.67**]	**0.42±0.34**[2.78]	94
10	0.97×0.97×0.97	1595	6.12±5.31[26.19]	0.52±0.79[24.11]	**0.40±0.42[6.81]**	93
Average	n/a	1427	6.41±5.11[22.78]	0.54±0.84[14.77]	**0.47±0.45[4.23]**	90

Digital phantoms have drawbacks when used to validate the accuracy of the proposed method. Xxx et al. reported that the respiratory motion calculated using DIR usually underestimates the true respiratory motion. Therefore, the digital phantom may not be adequate to simulate the real respiratory motion. The artificial DVF used to generate the moving image was generally very smooth. The results may be biased because the DIR algorithms including the PTVreg are tend to perform better when images to be registered are subject to smooth DVF. To avoid these drawbacks, we also used the real CT images and the 300 manual landmark pairs of the DIRLAB datasets to evaluate the accuracy of our method. Initial landmark pairs were established by propagating the 300 landmarks on the EI to the EE phase. Final landmarks pairs were established after the network refinement. Table 2 shows the TRE results, which is in good agreement with that in Table 1. Compared to the TREs of the initial landmark pairs, the TREs of the final landmark pairs were reduced from 0.83±0.87mm to 0.73±0.53. The average maximum TRE was almost halved after the network refinement. The TREs were smaller than 1mm for around 83% of the final landmark pairs. As of

the time of writing this paper, the average TREs for the finial landmark pairs (0.73±0.53mm) were more accurate than the average TREs of the 26 DIRs (0.91±1.07mm) that were posted at the DIRLAB website.

Table 2: TREs (mean±std[max]) of the proposed method on 10 DIRLAB datasets

Case #	**Voxel size (mm^3)**	**# of landmark pairs**	**Ground truth DVF (mm)**	**TRE for landmark pairs (mm)**		**% of final landmark pairs with TRE < 1mm**
				Initial	**Preserved**	
1	0.97×0.97×2.5	195	3.84±2.80[11.26]	0.70±0.60[2.70]	**0.69±0.43[2.29]**	83
2	1.16×1.16×2.5	198	4.36±3.82[17.54]	0.73±0.67[3.27]	**0.67±0.38[2.47]**	85
3	1.15×1.15×2.5	197	6.84±4.01[16.55]	0.80±0.72[5.21]	**0.74±0.47[2.91]**	80
4	1.13×1.13×2.5	197	9.82±4.89[19.57]	0.99±0.96[11.45]	**0.86±0.58[3.70]**	75
5	1.10×1.10×2.5	201	7.42±5.47[24.78]	0.90±1.27[15.95]	**0.73±0.64[7.91]**	83
6	0.97×0.97×2.5	201	10.88±6.98[27.24]	0.81±0.76[**4.63**]	**0.73±0.59**[4.83]	83
7	0.97×0.97×2.5	201	11.00±7.42[30.64]	0.79±0.68[5.71]	**0.73±0.59[5.67]**	86
8	0.97×0.97×2.5	198	15.00±9.00[30.45]	0.89±1.14[14.84]	**0.76±0.61[3.94]**	82
9	0.97×0.97×2.5	196	7.89±3.97[15.52]	0.84±0.76[**3.74**]	**0.77±0.65**[3.82]	81
10	0.97×0.97×2.5	202	7.36±6.40[27.79]	0.83±1.15[11.13]	**0.64±0.37[2.49]**	89
Average	n/a	199	8.44±5.47 [22.13]	0.83±0.87[7.86]	**0.73±0.53[4.00]**	83

3.2 The effectiveness of outlier rejection

Table 3 shows the comparison between TREs of the rejected and the preserved landmark pairs to demonstrate the effectiveness of the outlier rejection process. On average, 30% of the landmark pairs were rejected. The inaccurate network predictions with large maximum TREs were successfully rejected after the outlier rejection step. After the outlier rejection process, the average TRE was reduced from 0.96±1.52 to 0.47±0.45. The percentage of landmark pairs with TREs smaller than 1mm increased from 71% to 90% after the outlier rejection process. Figure 4 shows the positions of the preserved and rejected landmark pairs. The remaining landmark pairs after the rejection were uniformly positioned throughout the lung. A larger number of landmarks were rejected when the TRE were larger. This is because the MSPS network performed less robustly for landmark pairs with significant deformation.

Table 3: TREs (mean±std[max]) of the preserved and rejected landmark pairs on 10 digital phantoms

Case #	# of landmark pairs		TRE for landmark pairs (mm)		% of final landmark pairs with TRE < 1mm	
	Rejected	**Preserved**	**Rejected**	**Preserved**	**Rejected**	**Preserved**
1	574	1126	0.37±0.32[2.20]	0.34±0.29[1.67]	95	97
2	754	1546	0.71±0.76[3.99]	0.35±0.37[3.81]	77	95
3	599	1201	0.96±0.95[5.70]	0.51±0.55[4.76]	66	87
4	426	874	0.88±0.85[4.36]	0.55±0.54[4.86]	67	86
5	523	1077	1.07±1.12[6.11]	0.58±0.60[4.15]	64	83
6	590	1410	1.31±2.13[29.45]	0.57±0.55[4.39]	61	84
7	808	1792	1.12±2.65[29.99]	0.45±0.41[4.27]	68	92
8	1202	2798	1.79±4.58[40.72]	0.51±0.48[4.85]	59	87
9	445	855	0.52±0.44[2.31]	0.42±0.34[2.78]	86	94
10	705	1595	0.93±1.36[24.67]	0.40±0.42[6.81]	67	93
Average	662	1427	0.96±1.52[14.95]	0.47±0.45[4.23]	71	90

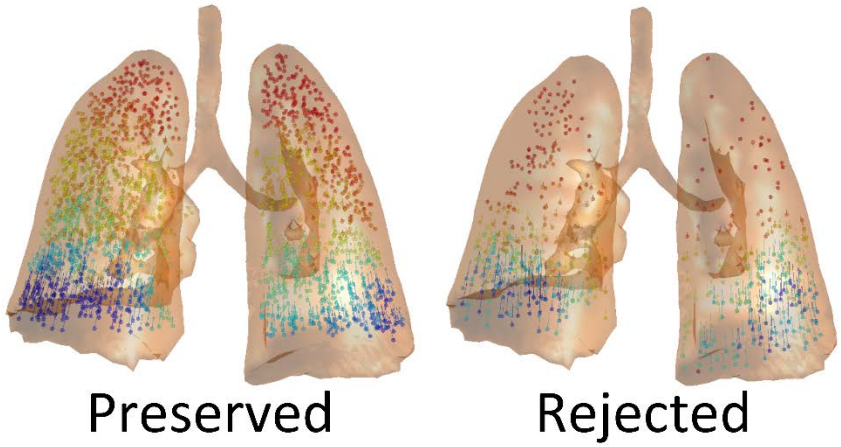

Fig. 4. The preserved landmark pairs and rejected landmark pairs for case 6

3.3 Our landmark pairs

a) Dense landmark pairs for DIRLAB datasets

Dense sets of landmark pairs were detected using the proposed method for the 10 DIRLAB 4DCT lung datasets. On average, 1886 landmark pairs were detected which is more than 6 times that of the 300 DIRLAB landmark pairs. To visualize the landmark pairs in 3D, the landmark pairs were plotted inside the

lung in Fig x for case 6, case 8 and case 9 in the side view, front view and top view respectively. DIRLAB 300 landmark pairs were also plotted for comparison.

Table 4: Number of landmark pairs detected for the 10 DIRLAB 4DCT lung datasets

Case #	1	2	3	4	5	6	7	8	9	10	average
# of Landmark pairs	1782	2235	1649	1276	1279	2072	2230	3121	1069	2151	1886

To assess the accuracy of the landmark pairs, manual blind spot checks were performed on 20 randomly selected landmark pairs for each case. Three observers were recruited and trained to perform the task. Given the landmarks on the EI phase, two postdocs manually placed the corresponding landmarks on the EE phase to generate two sets of manual landmark pairs. The two sets of manual landmark pairs were randomly shuffled with the automatically detected landmark pairs. The shuffled landmark pairs were ranked by the other observer without knowing how the landmark pair was produced. The result showed that the …

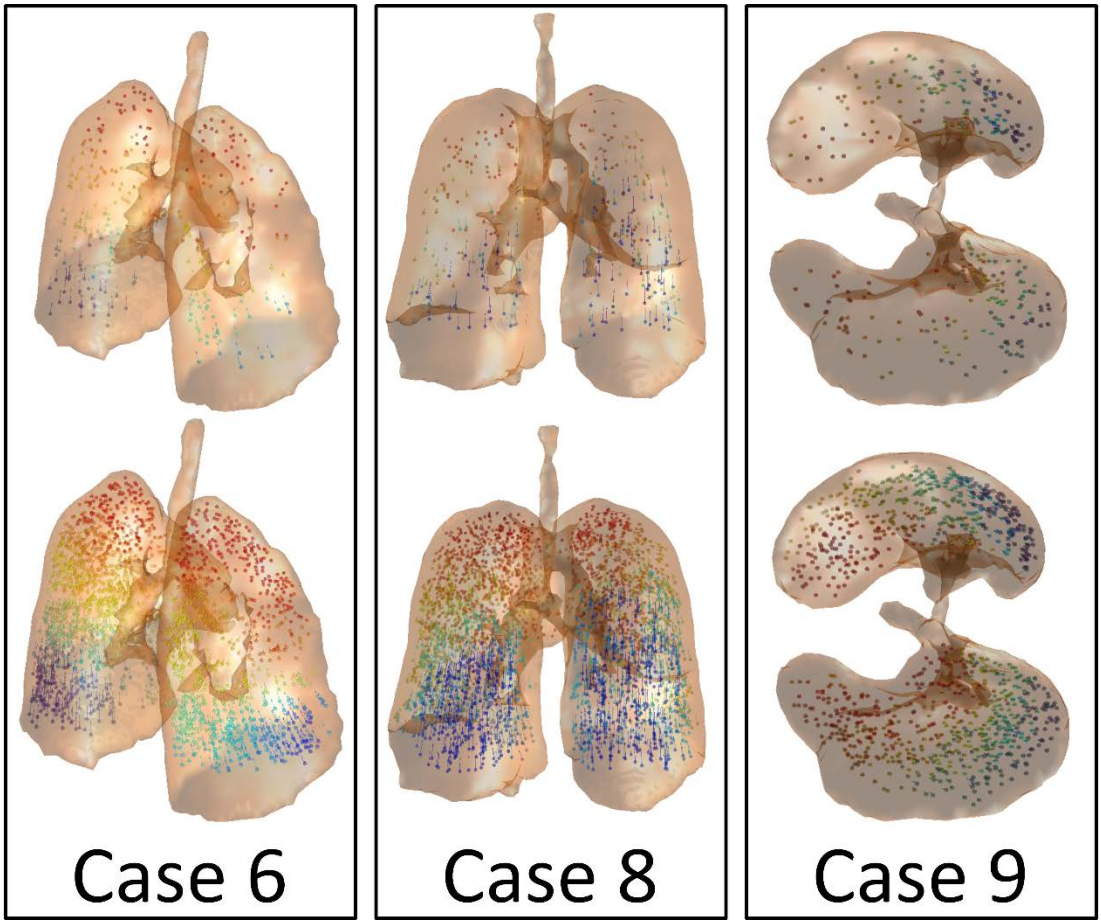

Fig. 5 Comparison of our dense set of landmark pairs (bottom row) with the DIRLAB 300 landmark pairs (top row), landmark pairs were color coded by its deformation vector magnitude.

b) Improved 300 manual landmark pairs of the DIRLAB datasets

Each of the 10 DIRLAB 4DCT landmark datasets has 300 manually labelled landmark pairs. The landmark labelling inter-observer uncertainties reported by the original authors were 0.88±1.31 mm on average which suggested that the quality of these landmark pairs could be further improved. The MSPS network was applied on the 300 manually placed landmark pairs to improve the landmark positional accuracy. A new set of landmarks were produced after the network refinement. To assess the positional accuracy, we asked 3 observer to manually verify the accuracy on selected landmark pairs with offset distance bigger than 1.7mm from its original positions. Only landmarks with offset distance bigger than 1.7 mm were used because it was easy to identify which landmark pair was better than the other when the offset distance is large. Table 5 shows that more landmarks were improved for cases with larger observers' uncertainties. Around 88% of the landmarks were actually improved after manual verification. An examples of landmark pair positional improvement post MSPS were shown in Fig.5. The position of the landmark at EE was shifted by a vector of (1.3, 0.9, 2.4) after MSPS refinement. It is obvious from the coronal and sagittal view that the landmark at EI was at the vessel corner and below the horizontal vessel. The landmark at EE was misplaced at location above the horizontal vessel prior to MSPS refinement. After the MSPS refinement, the landmark at EE was moved to the right location.

Table 5: Number of landmark pairs detected for the 10 DIRLAB 4DCT lung datasets

Cases	Observers Uncertainties mean±std (mm)	# of landmark pairs after refinement with offset distance > 1.7mm	# of landmark pairs that are improved after refinement		
			Observer 1	Observer 2	Observer 3
1	0.85±1.24	5	4		
2	0.70±0.99	4	3		
3	0.77±1.01	8	7		

4	1.13±1.27	29	27		
5	0.92±1.16	23	21		
6	0.97±1.38	11	9		
7	0.81±1.32	13	12		
8	1.03±2.19	18	17		
9	0.75±1.09	7	6		
10	0.86±1.45	28	23		
Total	n/a	146	129 (88.36%)		

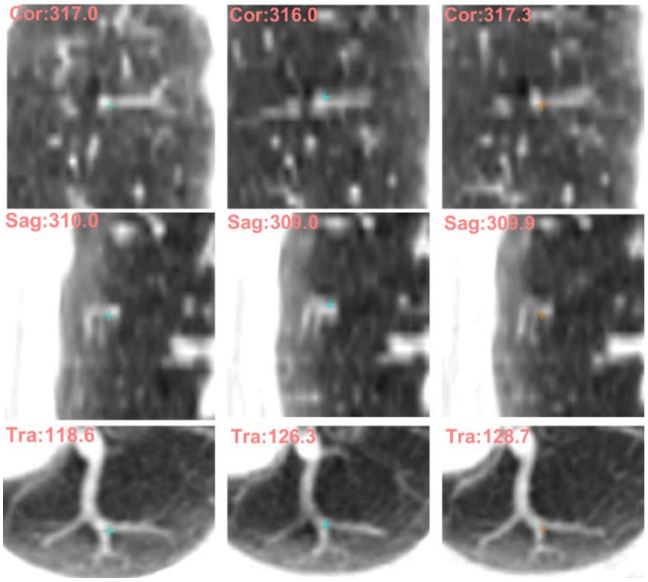

Fig. 4: Left column: landmark at EI, middle column: corresponding landmark at EE prior to MSPS refinement, right column: corresponding landmark at EE post MSPS refinement.

4 Discussion

The recently published AAPM task group report, TG132, outlined the important aspects of DIR for image guidance in radiotherapy. However, appropriate metrics were not yet defined for the evaluation of the geometric and dosimetric accuracy due to the lack of ground truth point to point correspondence. Paganelli et al reported an overview of patient-specific validation of DIR in radiation therapy, suggesting that accurate and efficient patient-specific validation was not available yet. Dense sets of anatomical landmarks could be used to drive the development of patient-specific DIR validation in radiotherapy applications. The proposed method can be used to validate DIR on not only the detected landmark pairs but also manually placed grid points within any region of interest. This can be achieved by applying the network

refinement on the corresponding image patches that are generated by the DIR to be evaluated. The resultant xyz vectors suggest how much the DIR is off from the 'ground truth'. Therefore, the proposed method could be used as not only a landmark pair generation tool but also a patient-specific QA tool.

The proposed method is a multi-step processes including landmark detection, initial correspondence establishment and positional accuracy refinement. One advantage is that these steps are independent from one another. New landmark detection methods with superior performance could be used to replace the current landmark detection method. We are currently developing a new method by detecting the bifurcations in 3D vasculature probability maps with thinned vessel trees. DIRs that are superior to the PTVreg can be used to establish the initial landmark correspondence. The MSPS network could also be replaced by another network with enhanced performance. The proposed network was trained on the 2700 DIRLAB landmark pairs which have very limited number of landmarks near the diaphragm where lung motion was significant. Hence, the network was less robust for landmarks near the diaphragm, leading to an increased number of predictions rejected near the diaphragm. This issue could be alleviated by detecting more landmarks and using more training datasets near the diaphragm. Future work to enhance the performance of the network includes 1) to test different network architectures with fine-tuned hyperparameters and 2) to increase the number of training datasets.

The network encoded the image patches to be compared into a common space by using the Siamese design with shared weights of layers. The idea could be extended to multimodality image patch comparison such as CT vs MRI, T1 vs T2 MRIs. However, the Siamese design is not recommended in this case since the two image patches to be compared are from different modalities. Layers with independent weights may be more efficient to encode the multimodality image patches into a common space for further comparison. Currently, a relatively simple thresholding method was used for the outlier rejection process. The performance of the outlier rejection process could be improved by using a simple classifier that is trained to differentiate the inaccurate predictions from the accurate predictions using the digital phantom datasets.

The detected landmark pairs could be used not only to validate a specific DIR but also to register images. The landmark pairs could be used as initial guidance or as additional constraints for intensity based DIRs [xx]. Deformation vector was calculated by using the landmark pairs near the landmark locations and by matching the image intensities away from the landmark locations. Image registration could also be performed using only the detected landmark pairs. One example is to use the moving least squares approximation, which computes a weighted least squares polynomial fit to the landmark pairs.

5 Conclusion

A new method was developed in this study to automatically and accurately detect large number of landmark pairs in 4DCT lung datasets for DIR validation. The dense landmark pairs were shared at www.xxx.com for future DIR validations.

Acknowledgement

The project described was partially supported by the AHRQ (Agency for Healthcare Research and Quality) grant number 1 R01 HS022888-01 and its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Agency for Healthcare Research and Quality.

Reference

1. Kadoya N. Use of deformable image registration for radiotherapy applications. *J Radiol Radiat Ther.* 2014;2.
2. Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: A survey. *IEEE transactions on medical imaging.* 2013;32(7):1153-1190.
3. Yang D, Brame S, El Naqa I, et al. DIRART -- A software suite for deformable image registration and adaptive radiotherapy research. *Medical Physics.* 2011;38(1):67-77.
4. Yan D, Vicini F, Wong J, Martinez A. Adaptive radiation therapy. *Physics in Medicine and Biology.* 1997;42(1):123-132.
5. Castadot P, Lee JA, Geets X, Grégoire V. Adaptive Radiotherapy of Head and Neck Cancer. *Seminars in radiation oncology.* 2010;20(2):84-93.
6. Hof H, Rhein B, Haering P, Kopp-Schneider A, Debus J, Herfarth K. 4D-CT-based target volume definition in stereotactic radiotherapy of lung tumours: Comparison with a conventional technique using individual margins. *Radiotherapy and Oncology.* 2009;93(3):419-423.
7. Chao M, Xie Y, Xing L. Auto-propagation of contours for adaptive prostate radiation therapy. *Physics in Medicine and Biology.* 2008;53(17):4533.
8. Faggiano E, Fiorino C, Scalco E, et al. An automatic contour propagation method to follow parotid gland deformation during head-and-neck cancer tomotherapy. *Physics in Medicine and Biology.* 2011;56(3):775.
9. Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Medical Physics.* 2014;41(5):050902.
10. Voet PWJ, Dirkx MLP, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJM. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. *Radiotherapy and Oncology.* 2011;98(3):373-377.
11. Li X, Wang X, Li Y, Zhang X. A 4D IMRT planning method using deformable image registration to improve normal tissue sparing with contemporary delivery techniques. *Radiation Oncology (London, England).* 2011;6:83-83.
12. Wang J, Gu X. High-quality four-dimensional cone-beam CT by deforming prior images. *Physics in Medicine and Biology.* 2013;58(2):231.

13. Wu G, Wang Q, Lian J, Shen D. Reconstruction of 4D-CT from a Single Free-Breathing 3D-CT by Spatial-Temporal Image Registration. *Information processing in medical imaging : proceedings of the conference.* 2011;22:686-698.

14. Cherpak A, Serban M, Seuntjens J, Cygler JE. 4D dose-position verification in radiation therapy using the RADPOS system in a deformable lung phantom. *Medical Physics.* 2011;38(1):179-187.

15. Niu CJ, Foltz WD, Velec M, Moseley JL, Al-Mayah A, Brock KK. A novel technique to enable experimental validation of deformable dose accumulation. *Medical Physics.* 2012;39(2):765-776.

16. Yeo UJ, Taylor ML, Supple JR, et al. Is it sensible to ``deform'' dose? 3D experimental validation of dose-warping. *Medical Physics.* 2012;39(8):5065-5072.

17. Castadot P, Geets X, Lee JA, Christian N, Grégoire V. Assessment by a deformable registration method of the volumetric and positional changes of target volumes and organs at risk in pharyngo-laryngeal tumors treated with concomitant chemo-radiation. *Radiotherapy and Oncology.* 2010;95(2):209-217.

18. Mencarelli A, van Kranen SR, Hamming-Vrieze O, et al. Deformable image registration for adaptive radiation therapy of head and neck cancer: accuracy and precision in the presence of tumor changes. *International journal of radiation oncology, biology, physics.* 2014;90(3):680-687.

19. Yang D, Zhang M, Chang X, et al. A method to detect landmark pairs accurately between intra-patient volumetric medical images. *Medical Physics.* 2017;44(11):5859-5872.

20. Lowe DG. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision.* 2004;60(2):91-110.

21. He X-C, Yung NH. Curvature scale space corner detector with adaptive threshold and dynamic region of support. Paper presented at: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on2004.

22. Morel J-M, Yu G. ASIFT: A new framework for fully affine invariant image comparison. *SIAM journal on imaging sciences.* 2009;2(2):438-469.

23. Mikolajczyk K, Tuytelaars T, Schmid C, et al. A comparison of affine region detectors. *International journal of computer vision.* 2005;65(1-2):43-72.

24. Mikolajczyk K, Schmid C. Scale & affine invariant interest point detectors. *International journal of computer vision.* 2004;60(1):63-86.

25. Oliveira FP, Tavares JMR. Medical image registration: a review. *Computer methods in biomechanics and biomedical engineering.* 2014;17(2):73-93.

26. Ourselin S, Roche A, Prima S, Ayache N. Block Matching: A General Framework to Improve Robustness of Rigid Registration of Medical Images. 2000; Berlin, Heidelberg.

27. Horn BK, Schunck BG. Determining optical flow. *Artificial intelligence.* 1981;17(1-3):185-203.

28. Maes F, Vandermeulen D, Suetens P. Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information. *Medical image analysis.* 1999;3(4):373-386.

29. Styner M, Brechbuhler C, Szckely G, Gerig G. Parametric estimate of intensity inhomogeneities applied to MRI. *IEEE transactions on medical imaging.* 2000;19(3):153-165.

30. Yang D, Li H, Low DA, Deasy JO, El Naqa I. A fast inverse consistent deformable image registration method based on symmetric optical flow computation. *Physics in Medicine and Biology.* 2008;53(21):6143-6165.

31. Wang H, Dong L, O'Daniel J, et al. Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy. *Physics in Medicine and Biology.* 2005;50(12):2887-2905.

32. Castillo R, Castillo E, Guerra R, et al. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Physics in Medicine and Biology*. 2009;54(7):1849.

33. Lindeberg T. Scale Selection Properties of Generalized Scale-Space Interest Point Detectors. *Journal of Mathematical Imaging and Vision*. 2013;46(2):177-210.