Physics in Medicine & Biology

PAPER

An unsupervised convolutional neural network-based algorithm for deformable image registration

To cite this article: Vasant Kearney *et al* 2018 *Phys. Med. Biol.* **63** 185017

IOP Publishing Phys. Med. Biol. 63 (2018) 185017 (10pp) https://doi.org/10.1088/1361-6560/aada66

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

An unsupervised convolutional neural network-based algorithm for deformable image registration

RECEIVED
18 March 2018

REVISED
17 July 2018

ACCEPTED FOR PUBLICATION
15 August 2018

PUBLISHED
17 September 2018

Vasant Kearney, Samuel Haaf, Atchar Sudhyadhom, Gilmer Valdes and Timothy D Solberg

Department of Radiation Oncology, University of California, San Francisco, CA, United States of America

E-mail: vasant.kearney@ucsf.edu

Keywords: deep learning, convolutional neural networks, cone beam CT, deformable image registration, unsupervised learning

Abstract

The purpose of the work is to develop a deep unsupervised learning strategy for cone-beam CT (CBCT) to CT deformable image registration (DIR). This technique uses a deep convolutional inverse graphics network (DCIGN) based DIR algorithm implemented on 2 Nvidia 1080 Ti graphics processing units. The model is comprised of an encoding and decoding stage. The fully-convolutional encoding stage learns hierarchical features and simultaneously forms an information bottleneck, while the decoding stage restores the original dimensionality of the input image. Activations from the encoding stage are used as the input channels to a sparse DIR algorithm. DCIGN was trained using a distributive learning-based convolutional neural network architecture and used 285 head and neck patients to train, validate, and test the algorithm. The accuracy of the DCIGN algorithm was evaluated on 100 synthetic cases and 12 hold out test patient cases. The results indicate that DCIGN performed better than rigid registration, intensity corrected Demons, and landmark-guided deformable image registration for all evaluation metrics. DCIGN required ~14 h to train, and ~3.5 s to make a prediction on a 512 × 512 × 120 voxel image. In conclusion, DCIGN is able to maintain high accuracy in the presence of CBCT noise contamination, while simultaneously preserving high computational efficiency.

1. Introduction

Image guided radiotherapy using cone beam computed tomography (CBCT) has become the standard of care in many clinics. However, widespread skepticism in regards to the accuracy and intelligibility of deformable image registration (DIR) has limited use of DIR (Pukala *et al* 2016). Intensity variation, noise contamination, and reconstruction artifacts in CBCT images makes CBCT to CT DIR particularly challenging. In an effort to circumvent the challenges of DIR, previous studies have focused on using CBCT images for direct dose computation (Yoo and Yin 2006, Yang *et al* 2007, Hatton *et al* 2009, Niu and Zhu 2011, Reeves *et al* 2012). However, poor CBCT image quality is subject to inaccurate Hounsfield Unit (HU) assignment, and thus, inaccurate dose calculation. Additionally, direct dose computation on CBCT necessitates re-segmentation of the targets and sensitive structures, placing further burden on clinical personnel. If the original simulation CT and corresponding contours could be deformably mapped to the CBCT, the dose could be calculated much more reliably, without increasing clinical workload (Chao *et al* 2008, Samant *et al* 2008, Xie *et al* 2008).

Although various scatter correction techniques have been proposed, non-linear intensity inconsistencies and scatter contamination remain an open problem in CBCT DIR (Rinkel *et al* 2007, Maltz *et al* 2008, Zhu *et al* 2009, Sun *et al* 2011). To overcome these challenges, many techniques have emerged that attempt to relate corresponding regions in images by some combination of intensity information. These methods implemented a wide range of gradient, frequency, or distribution information, and often use a pre-processing, post-processing, or iterative processing step to solve the CBCT to CT DIR problem (Guimond *et al* 2001, Lawson *et al* 2007, Hou *et al* 2011). However, all of these methods register voxels between the entire volumes of the corresponding image sets, which can lead to deformation inaccuracies, since not all regions within an image contain rich features.

Many intensity based feature selection algorithms that aim to distinguish a good set of candidate regions with correspondingly strong features have been implemented for medical image registration (Arsigny *et al* 2006, Ashburner 2007, Rueckert and Aljabar 2015). Features that correspond well with respect to intensity, however, do not necessarily correspond well in regards to anatomy (Rohlfing 2012, Wu *et al* 2012a). In an effort to mitigate the shortcomings of intensity based feature selection, many handcrafted feature selection algorithms have been implemented in DIR (Shen and Davatzikos 2002, Zhan and Shen 2006, Wu *et al* 2012b). However, because these feature selection methods are only designed for a specific imaging modality they do not generalize well (Wu *et al* 2006).

Alternative approaches utilizing supervised learning strategies have been the focus of recent efforts. One group of supervised learning strategies aims to directly predict the deformation between image sets (Ghosal and Ray 2017). This type of learning based strategy can be difficult, in regards to DIR, since ground truths are not always available. One method employed to address this is to create synthetic ground truths by applying known deformations to the image and then contaminating the image with noise and/or intensity transformations (Liao *et al* 2017). This approach can be useful as an evaluation metric when used in conjunction with other non-linear evaluation strategies (Shen *et al* 2017). However, it remains challenging to recreate certain clinical imaging characteristics that are present in CBCT images, such as metallic streak artifacts, motion artifacts, beam hardening artifacts, scatter, exponential edge gradient effects, and imaging noise. Furthermore, it may be difficult to accurately synthesize anatomical deformations that may be present between the CT and CBCT image sets. As an alternative to synthetic ground truths, some studies utilize expert-delineated ground truths between image sets. However, these data sets are cumbersome to create and are not always reliable, as intra-clinician variation between delineations can contaminate the dataset with human error (Wu *et al* 2007). Another group of supervised learning strategies focuses on selecting an optimal set of features from a large feature pool (Wu *et al* 2006, Ou *et al* 2011). The feature pool often includes a plethora of handcrafted expert-delineated features, but this strategy requires a large training data set with known correspondence between the set of images.

In response to the challenges associated with handcrafted and supervised learning strategies, Gaussian mixture models and K-means have been used to determine intrinsic features within images. These methods tend to suffer from the curse-of-dimensionality when the number of predictor variables is large, which can render K-means and Gaussian mixtures models cumbersome to use for high dimensional medical images (Mwangi *et al* 2014).

To overcome the shortcomings of naïve intrinsic feature extraction methods, unsupervised deep learning models have recently gained attention. Previous models utilizing auto-encoders have shown promise when registering mono-modality brain MRI images (Wu *et al* 2016). However, the MRI image sets used in these studies have similar image acquisition environments, and thus allow for shallower network design and less point matching discrimination than would be required in mixed modality image registration (Hu *et al* 2017, Balakrishnan *et al* 2018).

Recently, deep convolutional inverse graphics networks (DCIGN) used in computer science have been applied to solve imaging problems in the presence of noise contamination, intensity transformations, lighting and shading inconsistencies, and rotational variations (Kulkarni *et al* 2015). Historically, implementing DCIGNs to solve 3D medical imaging problems has been challenging, as conventional 3D DCIGNs require a large amount of memory. Recent advances and dramatic cost reduction in graphics processing units (GPUs), however, have made 3D DCIGN feasible.

In this study we take advantage of an adaptation of DCIGNs to improve the accuracy of CBCT to CT deformable image registration. The feasibility and accuracy of implementing a distributive learning based DCIGN model to deformably register CBCT to CT image sets is demonstrated for head and neck cancer patients undergoing radiation therapy.

2. Methods and materials

2.1. Deep convolutional inverse graphics networks

The network is comprised of two stages; a fully-convolutional encoding stage and a fully-connected decoding stage. The encoding stage learns hierarchical features and simultaneously forms an information bottleneck, while the decoding stage restores the original dimensionality of the input image. An overview of the architecture is provided in figure 1. Given a novel 3D input image, DCIGNs will predict a 4D set of hierarchical representations using a network of neurons with learnable biases and weights. To achieve this, 3 successive convolutional layers were implemented using 16 filters ($\Psi^{1,2,..,16}$), 12 filters ($\Omega^{1,2,..,12}$), and 8 filters ($\Theta^{1,2,..,8}$), respectively. Each convolutional layer utilized a convolutional kernel size of $5 \times 5 \times 5$, with each successive convolutional layer having a broader receptive field than the last. The receptive fields of the 1st, 2nd, and 3rd convolutional layers in the encoding stage are 5, 9, and 13 voxels respectively. A $3 \times 3 \times 3$ max pooling layer was used to downsample each of the 8 activations before they were passed into a fully connected layer $\Phi^{1000}_{1,2,..,8}$ that produced 1000 representations

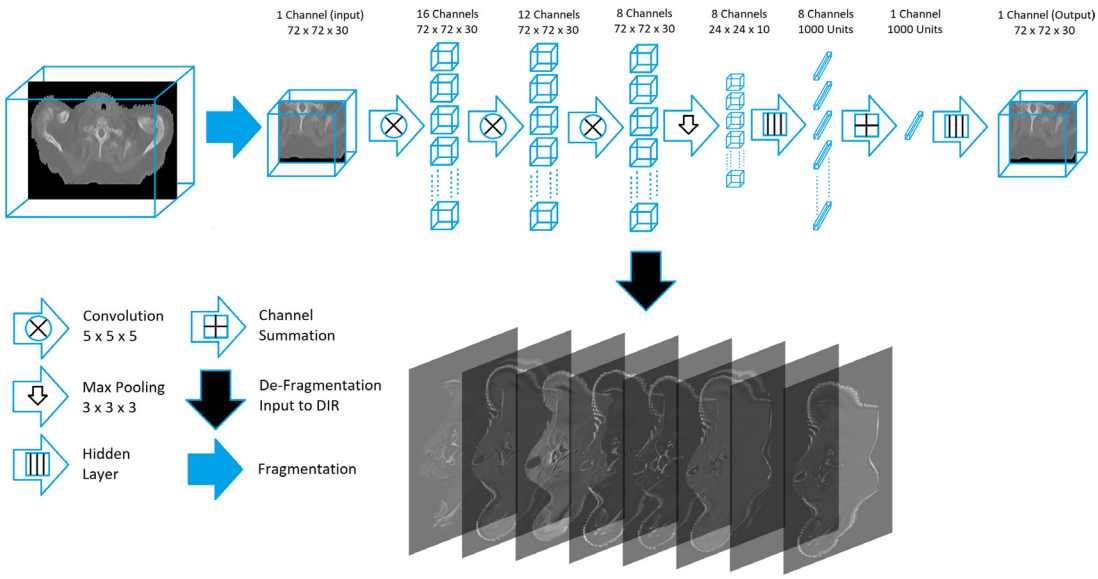

Figure 1. The model architecture is shown for the all steps within the network.

of each of the 8 downsampled filters. The summation of all 8 indices of $\Phi^{1000}_{1,2,\ldots,8}$ was taken to produce a bottleneck layer Φ^{1000}_{sum}. The combined representations of Φ^{1000}_{sum} were then upsampled using a fully connected layer Φ^{Final} and reshaped to match the dimensionality of the original input image. The root mean squared error of the linear output of the network and the original input image was used to compute the loss function. To help solve the vanishing gradient problem, each convolutional and fully connected layer used a rectified linear unit (ReLU) activation or an exponential linear unit (ELU) activation (LeCun *et al* 2015).

Due to memory limitations, all images were fragmented into block sizes of $72 \times 72 \times 30$ voxels, with a resolution of $1 \times 1 \times 3\,\text{mm}^3$. A zero padding of 4 was used for all convolutional layers. To mitigate edge effects, prediction blocks were overlapped and trimmed by 12, 12, and 6 voxels in the LR, AP, and SI directions respectively. All blocks from all training patients were shuffled, and trained in batch sizes of 4. A learning framework was implemented which distributed the model architecture and patient batches across 2 1080 Ti GTX Nvidia GPUs, having a total of 22 GB of GPU memory. The parameters from each half of the total batch were computed iteratively and have been implemented for medical image registration coalesced on a single GPU. The updated gradients were redistributed across all GPUs in an iterative fashion. Distributive learning allows for larger batch sizes and/or deeper model architectures.

To train the network, a first order stochastic gradient based optimizer, *Adam*, was used that adaptively estimates the lower-order moments (Kingma and Ba 2014). Adam performs automatic step size annealing by computing individual adaptive learning rates which are invariant to rescaling of the gradients. To mitigate overfitting, each convolutional and fully connected layer within the network used a dropout of 60% (Srivastava *et al* 2014). Batch normalization was used at each convolutional and fully connected layer to help with gradient saturation, and allow for reduced sensitivity to parameter initialization while enabling higher learning rates (Ioffe and Szegedy 2015).

Model generalization can be improved using data augmentation by encouraging the network to learn filters that are invariant to patient size, patient setup position, and patient posture. To help improve generalization, a data augmentation scheme was implemented using non-rigid deformations, rigid translation, contraction, and expansion.

In this study a total of 285 patient images were used: 221 for training and 40 for validation. An additional hold out set of 12 patients, comprised of 24 images, was used to report final deformation accuracy. During training, all images were rescaled to $1 \times 1 \times 3\,\text{mm}^3$, in the left-right, anterior–posterior, and superior–inferior directions, respectively, using trilinear interpolation. To train the model, a two-stage learning scheme was implemented. The first training stage used 3 data augmentations of each input image for a total of 663 images. 5000 randomly sampled fragmented $72 \times 72 \times 30$ voxel blocks were trained with 100 epochs, for a total of 125 000 iterations. The second training stage used the original 221 images, fragmented into 5000 patches and trained for another 100 epochs. This two stage approach helps the model find patient posture, size, and setup invariant filters in the first stage while simultaneously learning filters from the original decontaminated image set during the second learning stage. The model is trained on CT and CBCT images simultaneously, so that the filters can learn features that persist between imaging modalities. Figure 1 shows a schematic representation of the network architecture.

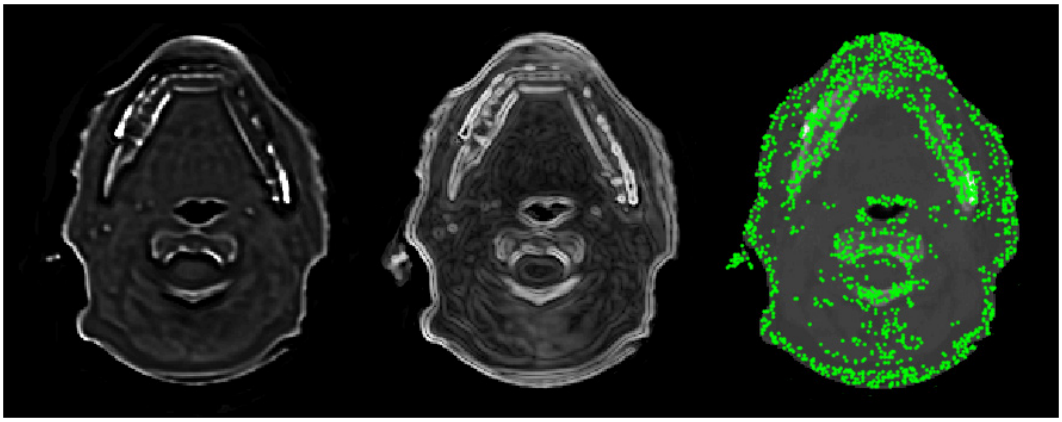

Figure 2. The mid-axial slice of the Θ^{5}_{moving} (left), $\nabla\Theta^{5}_{moving}$ (middle), and corresponding extracted points P^{5}_{Θ} superimposed on the original moving image (right).

2.2. Deformable image registration framework

Once the model was trained, the last two fully connected layers, $\Phi^{1000}_{1,2,..,8}$ and Φ^{Final}, were removed. This allows for the output of the Θ^{8} layer to feed directly into the deformable image registration framework (Forsberg *et al* 2011, Wu *et al* 2016). The 8 deep features $\Theta^{1,2,..,8}$ are used to find corresponding points between the filters of the CBCT ($\Theta^{1,2,..,8}_{static}$) and the CT ($\Theta^{1,2,..,8}_{moving}$) hold out image sets, where Θ_{static} and Θ_{moving} refer to the deep features associated with the CBCT and CT images, respectively. Not all regions of $\Theta^{1,2,..,8}_{moving}$ contain strong features, so an initial region discrimination step is implemented to identify a set of candidate points with strong deep features representations. To achieve this, the gradients of the feature representations $\nabla\Theta^{1,2,..,8}_{moving}$ are fed into a scale-invariant feature transformation (SIFT), which extracts a set of registration points ($P^{1,2,..,8}_{\Theta}$) (Lowe 2004). Figure 2 shows the point extraction process from the feature representations on the moving image.

A sparse multi-channel DIR algorithm was implemented to compute the deformation at every point (Forsberg *et al* 2011). Each feature representation was assigned its own channel such that r was computed in the following way;

$$\overset{\theta}{r}{}^{i}_{j} = \sum_{fn=1}^{8} \frac{\left(\theta^{\overset{i}{fn}}_{moving} - \theta^{fn}_{static}\right)\nabla\theta^{fn}_{static}}{\left(\theta^{\overset{i}{fn}}_{moving} - \theta^{fn}_{static}\right)^{2} + \left|\nabla\theta^{fn}_{static}\right|^{2}} + \frac{\left(\theta^{\overset{i}{fn}}_{moving} - \theta^{fn}_{static}\right)\nabla\theta^{\overset{i}{fn}}_{moving}}{\left(\theta^{\overset{i}{fn}}_{moving} - \theta^{fn}_{static}\right)^{2} + \left|\nabla\theta^{\overset{i}{fn}}_{moving}\right|^{2}}\Bigg|_{j}^{P^{1,2,..,8}_{\theta}}, \quad (1)$$

where $\overset{\theta}{r}{}^{i}_{j}$ represents the updated moving vector for each point j and iteration i, fn is the filter number, $\theta^{\overset{i}{fn}}_{moving}$ is the iteratively updated moving feature map, and θ^{fn}_{static} is the static feature map. θ is a box centered on each extracted point and locally confined to a $7 \times 7 \times 9\,\mathrm{mm}^3$ region. At every iteration, sparse interpolation was used to populate the full 3D deformation vector field (DVF) from the deformation of the extracted points r^{i}_{j}. The full iteratively updated image $\Theta^{\overset{i+1}{fn}}_{moving}$ was computed by deforming the moving $\Theta^{\overset{i}{fn}}_{moving}$ image with the accumulated vector field (Vercauteren *et al* 2009). The objective function was incrementally updated until the rate of convergence is below a threshold τ,

$$\text{where } \tau^{i} = \sum_{P^{1,2,..,8}_{\Theta}} \left|\overset{\Theta}{r}{}^{i+1}\right| \Big/ \sum_{P^{1,2,..,8}_{\Theta}} \left|\overset{\Theta}{r}{}^{i}\right|. \quad (2)$$

2.3. Evaluation

The accuracy of each DIR algorithm was evaluated by calculating the DVF error on a series of synthetic cases with known deformations. Synthetic data sets were generated by applying a series of 10 pre-determined DVFs, noise contaminations, and intensity transformations to 10 patient CT sets, producing a total of 100 ground truths. Previously reported electron density transformation values were used to rescale the intensity distribution of the image using a piecewise approach (Yoo and Yin 2006, Yang *et al* 2007, Hatton *et al* 2009, Reeves *et al* 2012). CBCT

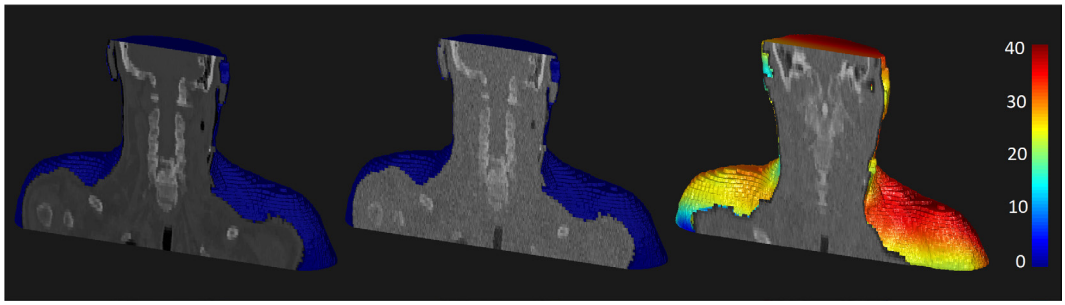

Figure 3. The body contour and coronal cross section of patient 2, can be seen for the undeformed image (left), undeformed image with noise contamination (center), and deformed image with noise contamination (right). Note that the image was contaminated with noise having ±200 HU and deformed with a maximum amplitude of 40 mm. The color map represents the magnitude of the deformation in millimeters.

were simulated by adding Poisson and Gaussian noise (Li *et al* 2004, Wang *et al* 2006). To mitigate the difference between the real and synthetic CBCT data sets, the noise, intensity transformation, and deformations were all computed randomly over a range of values. The DVFs were synthesized using a series of superimposed sinusoidal deformations with a maximum magnitude ranging from ~10 mm to ~40 mm. The standard deviation of the noise contamination ranged from ~0 to ~450 HU, with a mean HU contamination value of zero. Figure 3 shows the workflow of one synthetically generated test scenario.

Three different evaluation metrics were used to assess the accuracy of the patient cases: normalized mutual information (NMI), feature similarity index metric (FSIM), and root mean squared error of the 3D Canny edge ($RMSE_C$). NMI is a measurement of the corresponding intensity distributions between images, and is determined by computing the intensity transformation between the histograms from the moving image, static image, and joint histograms between the two images (Zhen *et al* 2012, Kearney *et al* 2017a). FSIM aims to emulate the human visual system by assessing the frequency and phase information between two corresponding image sets (Zhang *et al* 2011). The $RMSE_C$ between the Canny edges of the moving image C_{moving} and static image C_{static} is defined as $RMSE_C = \sum_i^N \min\left(\sqrt{\left(C^i_{moving} - C^j_{moving}\right)^2}\right)_j^M / N$, where N and M represent the length of all the points from the moving and static images respectively (Kearney *et al* 2017b). The NMI and FSIM values were normalized from 0 to 1, where 1 represents maximum similarity between the image sets.

Twelve patients were used to test the null hypothesis that DCIGN and the other DIR methods are from populations with equal $RMSE_C$, FSIM, NMI means. A two-sample one-sided Mann–Whitney U test was used, since it does not assume a Gaussian distribution. A p-value of 0.05 was used to reject the null hypothesis.

To test the validity of a Gaussian distribution, an Anderson–Darling goodness-of-fit hypothesis test was used. A Mann–Whitney U test was used instead of a Student t-test, since the Anderson–Darling test determined that a Gaussian assumption was not valid for all evaluation metrics.

3. Results

3.1. Synthetic performance

Intensity corrected Demons (IC Demons), landmark-guided deformable image registration (LDIR), and DCIGN were tested using 100 different synthetic scenarios (Zhen *et al* 2012, Kearney *et al* 2014). Figure 4 shows the results for all synthetic test cases for all three DIR algorithms. A linear regression slope is shown for the 95% error margin versus intensity noise for all algorithms.

Figure 5 shows a comparison between the various algorithms for a synthetic case with a standard deviation of 142 HU, and a maximum deformation of 21.5 mm. The magnitude of the error between the DVF from each deformation algorithm and the ground DVF are shown with a heat map superimposed on the CT image.

3.2. Performance on patient data

The NMI, FSIM, and the $RMSE_C$ values are shown in tables 1–3 respectively for all 12 patients using all DIR algorithms and rigid registration (Zhen *et al* 2012, Kearney *et al* 2014). For NMI and FSIM a higher value indicates better deformation congruence. For $RMSE_C$ lower values indicate better agreement between the Canny edges of the two corresponding images. Table 4 presents the mean and standard deviation for all evaluation metrics on all algorithms as well as the corresponding p-values. The corresponding p-values for the Mann–Whitney U test are also shown.

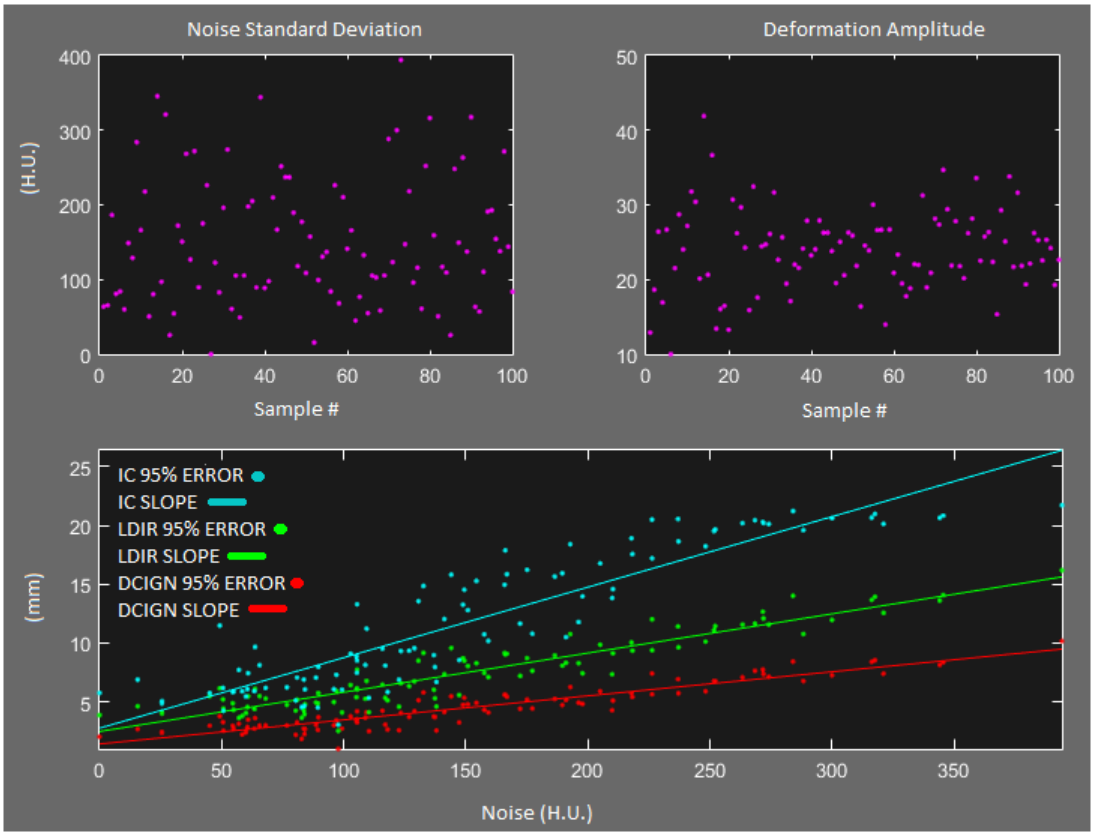

Figure 4. The standard deviation (top-left) and the corresponding maximum deformation amplitude (top-right) are shown for each of the 100 samples. The 95% error margin for the IC Demons, LDIR and DCIGN methods are shown versus intensity noise for all samples, with the corresponding linear regression slope (bottom).

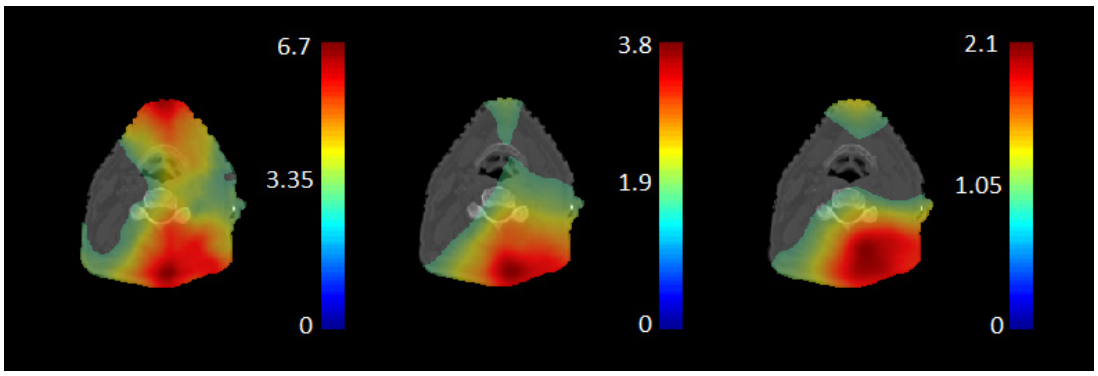

Figure 5. The magnitude of the deformation error for the ground truth is superimposed on a mid-axial slice for the IC Demons (left), LDIR (center), and DCIGN (right). The transparency of the color map is scaled linearly with the magnitude of the deformation error and is thresholded at 20% of the maximum error. The color map represents the magnitude of the error in millimeters.

The DCIGN algorithm was able to achieve higher NMI and FSIM than all other DIR algorithms for all 12 patients. On average, the DCIGN produced closer Canny edges (lower $RMSE_C$) between the corresponding images than all other algorithms for all 12 patients. To help demonstrate the DVF characteristics of each algorithm, a deformation grid superimposed on patient 1 is shown in figure 6. Tables 1–3 indicate that patient 1 achieved better deformation accuracy with DCIGN than IC Demons, or LDIR, which can be qualitatively observed in figure 6 by the deformation grids superimposed on the CBCT image.

3.3. Learning and prediction workflow

Since data augmentation is imperfect and introduces sampling errors, a two-stage training scheme was implemented. During the first stage the model trains only on the augmented data (first 100 epochs) and during the second stage the model trains only on the original data (last 100 epochs). Figure 7 shows the loss value for

Table 1. The NMI values are shown table for all 12 patients algorithms.

		P1	P2	P3	P4	P5	P6
NMI	Rigid registration	0.59	0.52	0.49	0.6	0.59	0.53
	IC Demons	0.61	0.59	0.54	0.64	0.65	0.58
	LDIR	0.66	0.59	0.55	0.66	0.66	0.58
	DCIGN	0.69	0.62	0.56	0.68	0.68	0.63
		P7	P8	P9	P10	P11	P12
	Rigid registration	0.51	0.53	0.54	0.58	0.65	0.64
	IC Demons	0.58	0.58	0.57	0.63	0.71	0.7
	LDIR	0.58	0.57	0.6	0.65	0.73	0.69
	DCIGN	0.61	0.62	0.63	0.66	0.74	0.72

Table 2. The FSIM values are shown for all 12 patients for all algorithms.

		P1	P2	P3	P4	P5	P6
FSIM	Rigid registration	0.82	0.8	0.79	0.87	0.86	0.84
	IC Demons	0.83	0.8	0.82	0.89	0.85	0.87
	LDIR	0.87	0.84	0.81	0.91	0.91	0.87
	DCIGN	0.9	0.86	0.86	0.94	0.93	0.91
		P7	P8	P9	P10	P11	P12
	Rigid registration	0.85	0.85	0.86	0.85	0.89	0.9
	IC Demons	0.86	0.85	0.88	0.89	0.92	0.92
	LDIR	0.88	0.87	0.91	0.9	0.93	0.94
	DCIGN	0.92	0.91	0.94	0.93	0.96	0.97

Table 3. The $RMSE_C$ values are shown table for all 12 patients for all algorithms.

		P1	P2	P3	P4	P5	P6
$RMSE_C$	Rigid registration	0.29	0.27	0.25	0.23	0.23	0.25
	IC Demons	0.27	0.26	0.23	0.21	0.21	0.21
	LDIR	0.25	0.23	0.21	0.18	0.18	0.21
	DCIGN	0.21	0.2	0.19	0.16	0.16	0.18
		P7	P8	P9	P10	P11	P12
	Rigid registration	0.23	0.25	0.26	0.29	0.2	0.22
	IC Demons	0.2	0.23	0.22	0.26	0.18	0.19
	LDIR	0.18	0.21	0.21	0.25	0.14	0.16
	DCIGN	0.16	0.18	0.17	0.21	0.13	0.15

Table 4. The mean and standard deviation are shown for all evaluation metrics with all algorithms as well as the corresponding *p*-values.

	NMI				FSIM				$RMSE_C$			
	Mean	±	STDEV	*P*-value	Mean	±	STDEV	*P*-value	Mean	±	STDEV	*P*-value
Rigid registration	0.564	±	0.052	<0.001	0.848	±	0.033	<0.001	0.248	±	0.027	<0.001
IC Demons	0.615	±	0.052	0.042	0.865	±	0.038	<0.001	0.223	±	0.029	<0.001
LDIR	0.627	±	0.056	0.117	0.887	±	0.037	0.019	0.201	±	0.034	0.022
DCIGN	0.653	±	0.051		0.919	±	0.034		0.175	±	0.025	

the two-stage learning model. Each iteration represents 1250 iterations. The validation loss and training loss are reported at every epoch to reduce computation time.

The two-stage learning scheme required ~14 h to train 200 epochs on two GPUs using a batch size of 4 with a total of 5000 patches. In contrast, it takes ~3.5 s to fragment an image of size 512 × 512 × 120 voxels image into 72 × 72 × 30 voxel blocks, predict on all blocks, and defragment the 8 deep hierarchal representative blocks into a 512 × 512 × 120 × 8 4D matrix. The validation and training losses set tended to diverge after roughly 100 epochs, so each leg of the two-stage learning scheme was stopped at 100 epochs.

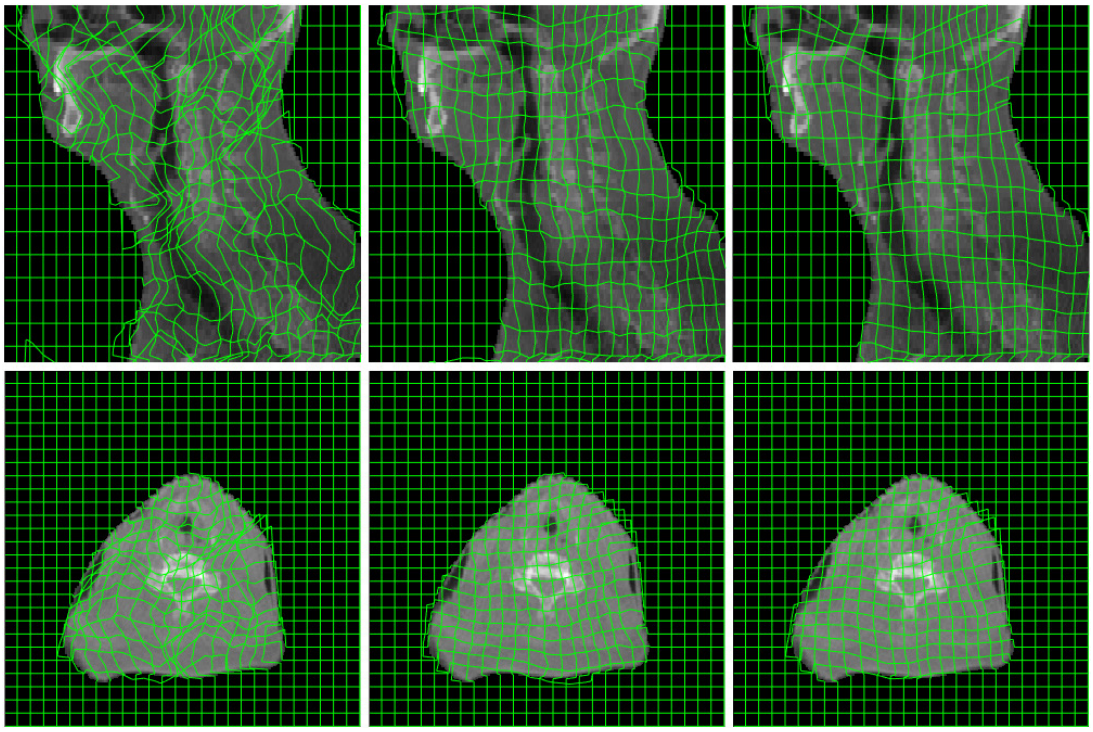

Figure 6. A deformed grid superimposed of the static image is shown for the IC Demons (left), LDIR (center), and DCIGN (right).

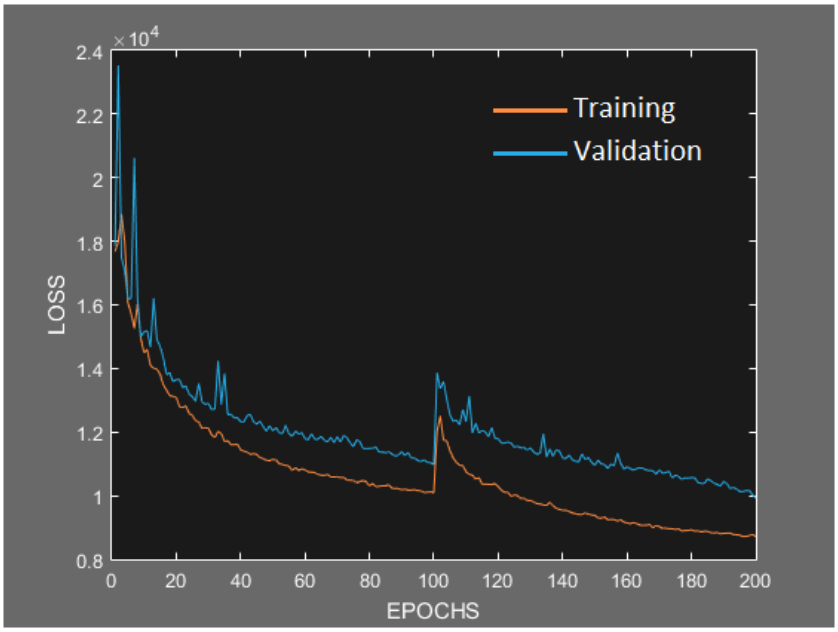

Figure 7. The loss is shown for the validation and training data sets for 5000 fragmented images with a batch size of 4.

4. Discussion

The scalability of the various DIR algorithms was demonstrated using the synthetic cases, in which the IC Demons algorithm performed similar to LDIR and DCIGN with noise contamination below ±30 HU at the 95% error margin. With noise contamination below ±100 HU, the DCIGN algorithm performed significantly better than all other algorithms at the 95% error margin. The IC Demons algorithm had the steepest linear regression slope, which indicates that it behaves worse than the other algorithms as noise levels increase. LDIR and IC Demons had a similar linear regression intercept. DCIGN had a slightly lower linear regression intercept than LDIR and IC Demons, which indicates that it performed slightly better at low noise levels. The DCIGN algorithm had a much

more shallow slope than either LDIR or IC Demons, indicating that it behaves well in the presence of large noise contaminations.

While the NMI mean values for the DCIGN algorithm were the highest, this did not represent a statistically significant improvement over the LDIR algorithm. NMI characterizes the joint intensity distribution between two image sets and does not always account for inaccurate deformations between two image sets. Non-physical deformations can improve the NMI values while simultaneously degrading the accuracy of the DVF. The FSIM values were statistically better for the DCIGN algorithm than all other algorithms. Similarly, DCIGN outperformed all other algorithms for the $RMSE_C$ evaluation metric and had a statistically significant improvement over all other methods.

The depth and architecture of the network was designed to stay within the memory limitations of the hardware used. Down-sampling via convolution with corresponding de-convolutions was not used in order to maintain the original resolution of the input image. Fully connected layers have the advantage of large receptive fields and tend to converge well, but also suffer from over-fitting. To improve generalization, small batch sizes, dropout, data augmentation, two-stage model training, and a low number of filters was used.

Even with a 2-GPU distributed learning architecture, the batch size was limited to only 4 image blocks per iteration. It is possible that the deep learning model would have converged more accurately if larger batch sizes were used. Similarly, the model depth and the number of filters could have been increased if more GPUs were used.

Future projects are planned with additional GPUs, which will implement deeper architectures with more filters. It is anticipated that these models will improve the accuracy of DCIGN, but will be also at greater risk to over-fitting. Since the number of parameters will increase, different training techniques, such as L1 or L2 regularization, additional dropout, additional data augmentation, leaky rectifying linear units, or parametric rectifying linear units, will have to be used to improve generalization. Finally, these algorithms will be implemented on a broader range of imaging data, including publically available imaging sets, which will help the readership benchmark performance.

5. Conclusions

Given sufficient training data, the DCIGN algorithm outperformed all other algorithms for the synthetic and patient cases. By moving towards a deep learning based unsupervised learning strategy, CBCT-CT DIR can be accurate in the presence of heavy noise contamination, while simultaneously preserving high computational efficiency.

References

Arsigny V, Commowick O, Pennec X and Ayache N 2006 A log-euclidean framework for statistics on diffeomorphisms *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2006* pp 924–31

Ashburner J 2007 A fast diffeomorphic image registration algorithm *NeuroImage* **38** 95–113

Balakrishnan G, Zhao A, Sabuncu M R, Guttag J and Dalca A V 2018 An unsupervised learning model for deformable medical image registration (arXiv:1802.02604)

Chao M, Xie Y and Xing L 2008 Auto-propagation of contours for adaptive prostate radiation therapy *Phys. Med. Biol.* **53** 4533–42

Forsberg D, Rathi Y, Bouix S, Wassermann D, Knutsson H and Westin C-F 2011 Improving registration using multi-channel diffeomorphic demons combined with certainty maps *Multimodal Brain Image Analysis* (Berlin: Springer) pp 19–26

Ghosal S and Ray N 2017 Deep deformable registration: enhancing accuracy by fully convolutional neural net *Pattern Recognit. Lett.* **94** 81–6

Guimond A, Roche A, Ayache N and Meunier J 2001 Three-dimensional multimodal brain warping using the demons algorithm and adaptive intensity corrections *IEEE Trans. Med. Imaging* **20** 58–69

Hatton J, McCurdy B and Greer P B 2009 Cone beam computerized tomography: the effect of calibration of the Hounsfield unit number to electron density on dose calculation accuracy for adaptive radiation therapy *Phys. Med. Biol.* **54** N329–46

Hou J, Guerrero M, Chen W and D'Souza W D 2011 Deformable planning CT to cone-beam CT image registration in head-and-neck cancer *Med. Phys.* **38** 2088

Hu S, Wei L, Gao Y, Guo Y, Wu G and Shen D 2017 Learning-based deformable image registration for infant MR images in the first year of life *Med. Phys.* **44** 158–70

Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift (arXiv:1502.03167)

Kearney V, Chen S, Gu X, Chiu T, Liu H, Jiang L, Wang J, Yordy J, Nedzi L and Mao W 2014 Automated landmark-guided deformable image registration *Phys. Med. Biol.* **60** 101

Kearney V, Cheung J P, McGuinness C and Solberg T D 2017a CyberArc: a non-coplanar-arc optimization algorithm for CyberKnife *Phys. Med. Biol.* **62** 5777

Kearney V, Huang Y, Mao W, Yuan B and Tang L 2017b Canny edge-based deformable image registration *Phys. Med. Biol.* **62** 966

Kingma D and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)

Kulkarni T D, Whitney W F, Kohli P and Tenenbaum J 2015 Deep convolutional inverse graphics network *Advances in Neural Information Processing Systems* pp 2539–47

Lawson J D, Schreibmann E, Jani A B and Fox T 2007 Quantitative evaluation of a cone-beam computed tomography-planning computed tomography deformable image registration method for adaptive radiation therapy *J. Appl. Clin. Med. Phys.* **8** 2432

LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436
Li T, Li X, Wang J, Wen J, Lu H, Hsieh J and Liang Z 2004 Nonlinear sinogram smoothing for low-dose x-ray CT *IEEE Trans. Nucl. Sci.* **51** 2505–13
Liao R, Miao S, de Tournemire P, Grbic S, Kamen A, Mansi T and Comaniciu D 2017 An artificial agent for robust image registration *AAAI* pp 4168–75
Lowe D G 2004 Distinctive image features from scale-invariant keypoints *Int. J. Comput. Vis.* **60** 91–110
Maltz J S, Gangadharan B, Bose S, Hristov D H, Faddegon B A, Paidi A and Bani-Hashemi A R 2008 Algorithm for x-ray scatter, beam-hardening, and beam profile correction in diagnostic (kilovoltage) and treatment (megavoltage) cone beam CT *IEEE Trans. Med. Imaging* **27** 1791–810
Mwangi B, Tian T S and Soares J C 2014 A review of feature reduction techniques in neuroimaging *Neuroinformatics* **12** 229–44
Niu T and Zhu L 2011 Scatter correction for full-fan volumetric CT using a stationary beam blocker in a single full scan *Med. Phys.* **38** 6027–38
Ou Y, Sotiras A, Paragios N and Davatzikos C 2011 DRAMMS: deformable registration via attribute matching and mutual-saliency weighting *Med. Image Anal.* **15** 622–39
Pukala J, Johnson P B, Shah A P, Langen K M, Bova F J, Staton R J, Mañon R R, Kelly P and Meeks S L 2016 Benchmarking of five commercial deformable image registration algorithms for head and neck patients *J. Appl. Clin. Med. Phys.* **17** 25–40
Reeves T E, Mah P and McDavid W D 2012 Deriving Hounsfield units using grey levels in cone beam CT: a clinical application *Dentomaxillofacial Radiol.* **41** 500–8
Rinkel J, Gerfault L, Esteve F and Dinten J M 2007 A new method for x-ray scatter correction: first assessment on a cone-beam CT experimental setup *Phys. Med. Biol.* **52** 4633–52
Rohlfing T 2012 Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable *IEEE Trans. Med. Imaging* **31** 153–63
Rueckert D and Aljabar P 2015 *Handbook of Biomedical Imaging* (Berlin: Springer) pp 277–94
Samant S S, Xia J, Muyan-Ozcelik P and Owens J D 2008 High performance computing for deformable image registration: towards a new paradigm in adaptive radiotherapy *Med. Phys.* **35** 3546–53
Shen D and Davatzikos C 2002 HAMMER: hierarchical attribute matching mechanism for elastic registration *IEEE Trans. Med. Imaging* **21** 1421–39
Shen D, Wu G and Suk H-I 2017 Deep learning in medical image analysis *Annu. Rev. Biomed. Eng.* **19** 221–48
Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58
Sun M, Nagy T, Virshup G, Partain L, Oelhafen M and Star-Lack J 2011 Correction for patient table-induced scattered radiation in cone-beam computed tomography (CBCT) *Med. Phys.* **38** 2058–73
Vercauteren T, Pennec X, Perchant A and Ayache N 2009 Diffeomorphic demons: efficient non-parametric image registration *NeuroImage* **45** S61–72
Wang J, Li T, Lu H and Liang Z 2006 Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose x-ray computed tomography *IEEE Trans. Med. Imaging* **25** 1272–83
Wu G, Kim M, Wang Q, Munsell B C and Shen D 2016 Scalable high-performance image registration framework by unsupervised deep feature representations learning *IEEE Trans. Biomed. Eng.* **63** 1505–16
Wu G, Kim M, Wang Q and Shen D 2012a Hierarchical attribute-guided symmetric diffeomorphic registration for MR brain images *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Berlin: Springer) pp 90–7
Wu G, Qi F and Shen D 2006 Learning-based deformable registration of MR brain images *IEEE Trans. Med. Imaging* **25** 1145–57
Wu G, Qi F and Shen D 2007 Learning best features and deformation statistics for hierarchical registration of MR brain images *Information Processing in Medical Imaging* (Berlin: Springer) pp 160–71
Wu G, Wang Q, Jia H and Shen D 2012b Feature-based groupwise registration by hierarchical anatomical correspondence detection *Hum. Brain Mapp.* **33** 253–71
Xie Y, Chao M, Lee P and Xing L 2008 Feature-based rectal contour propagation from planning CT to cone beam CT *Med. Phys.* **35** 4450–9
Yang Y, Schreibmann E, Li T, Wang C and Xing L 2007 Evaluation of on-board kV cone beam CT (CBCT)-based dose calculation *Phys. Med. Biol.* **52** 685–705
Yoo S and Yin F-F 2006 Dosimetric feasibility of cone-beam CT-based treatment planning compared to CT-based treatment planning *Int. J. Radiat. Oncol. Biol. Phys.* **66** 1553–61
Zhan Y and Shen D 2006 Deformable segmentation of 3D ultrasound prostate images using statistical texture matching method *IEEE Trans. Med. Imaging* **25** 256–72
Zhang L, Zhang L, Mou X and Zhang D 2011 FSIM: a feature similarity index for image quality assessment *IEEE Trans. Image Process.* **20** 2378–86
Zhen X, Gu X, Yan H, Zhou L, Jia X and Jiang S B 2012 CT to cone-beam CT deformable registration with simultaneous intensity correction *Phys. Med. Biol.* **57** 6807
Zhu L, Xie Y, Wang J and Xing L 2009 Scatter correction for cone-beam CT in radiation therapy *Med. Phys.* **36** 2258–68