IOP Publishing Phys. Med. Biol. 65 (2020) 015011 (13pp) https://doi.org/10.1088/1361-6560/ab5da0

Physics in Medicine & Biology

Institute of Physics and Engineering in Medicine

RECEIVED
17 August 2019

REVISED
7 November 2019

ACCEPTED FOR PUBLICATION
29 November 2019

PUBLISHED
13 January 2020

PAPER

A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration

Zhuoran Jiang[1,2], **Fang-Fang Yin**[2,3,4], **Yun Ge**[1] **and Lei Ren**[2,3,5]

[1] School of Electronic Science and Engineering, Nanjing University, 163 Xianlin Road, Nanjing, Jiangsu 210046, People's Republic of China
[2] Department of Radiation Oncology, Duke University Medical Center, DUMC Box 3295, Durham, NC 27710, United States of America
[3] Medical Physics Graduate Program, Duke University, 2424 Erwin Road Suite 101, Durham, NC 27705, United States of America
[4] Medical Physics Graduate Program, Duke Kunshan University, Kunshan, Jiangsu 215316, People's Republic of China
[5] Author to whom any correspondence should be addressed.

E-mail: lei.ren@duke.edu

Keywords: deformable image registration, multi-scale framework, joint training, convolutional neural network, unsupervised learning

Abstract

To achieve accurate and fast deformable image registration (DIR) for pulmonary CT, we proposed a Multi-scale DIR framework with unsupervised Joint training of Convolutional Neural Network (MJ-CNN). MJ-CNN contains three models at multi-scale levels for a coarse-to-fine DIR to avoid being trapped in a local minimum. It is trained based on image similarity and deformation vector field (DVF) smoothness, requiring no supervision of ground-truth DVF. The three models are first trained sequentially and separately for their own registration tasks, and then are trained jointly for an end-to-end optimization under the multi-scale framework. In this study, MJ-CNN was trained using public SPARE 4D-CT data. The trained MJ-CNN was then evaluated on public DIR-LAB 4D-CT dataset as well as clinical CT-to-CBCT and CBCT-to-CBCT registration. For 4D-CT inter-phase registration, MJ-CNN achieved comparable accuracy to conventional iteration optimization-based methods, and showed the smallest registration errors compared to recently published deep learning-based DIR methods, demonstrating the efficacy of the proposed multi-scale joint training scheme. Besides, MJ-CNN trained using one dataset (SPARE) could generalize to a different dataset (DIR-LAB) acquired by different scanners and imaging protocols. Furthermore, MJ-CNN trained on 4D-CTs also performed well on CT-to-CBCT and CBCT-to-CBCT registration without any re-training or fine-tuning, demonstrating MJ-CNN's robustness against applications and imaging techniques. MJ-CNN took about 1.4 s for DVF estimation and required no manual-tuning of parameters during the evaluation. MJ-CNN is able to perform accurate DIR for pulmonary CT with nearly real-time speed, making it very applicable for clinical tasks.

1. Introduction

Deformable image registration (DIR) plays an important role in medical image analysis, which aligns paired images by establishing accurate dense deformation vector field (DVF). Conventional DIR iteratively optimizes the transformation model parameters to minimize predefined dissimilarity metrics and to enforce the DVF smoothness, which is technically an optimization problem. The registration process usually requires intensive computing and intricate parameter-tuning for the testing cases, making it time-consuming and user-dependent.

In recent years, deep learning-based methods have been introduced to address the limitations of conventional methods. Those methods can be divided into two categories: (1) integrated deep learning (Wu *et al* 2016, Yang *et al* 2017, Kearney *et al* 2018), and (2) end-to-end deep learning (Cao *et al* 2017, Eppenhof and Pluim 2017, Sokooti *et al* 2017, Shan *et al* 2017, Balakrishnan *et al* 2018, de Vos *et al* 2019, Eppenhof and Pluim 2018a).

For the first category, deep learning-based methods have been integrated into the conventional iterative optimization-based DIR methods to improve their performance. Wu *et al* (2016) proposed a convolutional stacked autoencoder to learn the low-dimensional feature representation, which was integrated into the conventional registration methods to improve the DIR accuracy for brain magnetic resonance imaging (MRI). For the head and neck CBCT to CT registration, Kearney *et al* (2018) used a deep convolution inverse graphic networks to learn the feature representation and fed the features into the DIR framework. Yang *et al* (2017) introduced a fast DIR method based on an encoder-decoder model to predict the large deformation diffeomorphic metric mapping (LDDMM) momentum-parameterization for brain MRI registration.

For the second category, deep learning-based methods have been developed to completely replace the conventional iterative optimization-based methods to achieve an accurate and fast DVF prediction. These methods can be classified into supervised learning and unsupervised learning.

The supervised learning-based methods train the models to minimize the difference between the predicted DVF and the ground truth DVF (Eppenhof and Pluim 2017, 2018b, Cao *et al* 2017, Sokooti *et al* 2017). One challenge for supervised learning is that the ground truth DVF is hard to obtain for real patients. A typical solution is to train the model on synthetic deformations (Sokooti *et al* 2017, Eppenhof and Pluim 2018a). For example, a recent study reported by Eppenhof *et al* (Eppenhof and Pluim 2018b) introduced a three-dimensional (3D) U-Net to directly predict DVF from the input paired volumes, where the ground truth was the synthetic random transformations. Performance of such methods is highly dependent on the accuracy of the ground truth DVF synthesized. In addition, it is challenging to generate realistic synthetic deformation for different anatomical sites.

Compared to the supervised learning-based methods, unsupervised learning-based methods do not need ground truth DVF during the training process. Instead, these methods train the models to minimize the dissimilarity between the deformed source image and the target image and to penalize the local spatial variations in the DVF domain. An unsupervised model for DIR consists of two parts: a convolutional neural network (CNN) for feature extraction and DVF estimation, and a spatial transformer layer for warping the source images to match the target images. Shan *et al* (2017) proposed an unsupervised end-to-end strategy for 2D CT/MRI registration by estimating dense DVF between paired images using CNN. However, deformations in medical images are usually not limited to 2D slices. To explore the 3D context, Balakrishnan *et al* (2018) introduced an unsupervised learning-based solution (VoxelMorph) for 3D MRI DIR, in which the CNN model was trained for the dense DVF estimation. Yet their network was designed on a single scale. To register large and complex deformations, the multi-scale scheme is generally needed to avoid being trapped in a local minimum. Recently, a deep learning image registration (DLIR) framework was proposed by de Vos *et al* (2019) for medical image registration using a multi-scale strategy. In their work, CNN models were trained for either affine image registration or DIR based on B-Spline. By stacking multiple models at multi-scale levels, a coarse-to-fine DIR was performed. The multi-scale framework used in the DLIR achieved fast and decent registration for various sites. However, as shown in their results of pulmonary 4D-CT registration, the DLIR method had large registration errors for cases with large deformations.

In this study, we proposed a Multi-scale DIR framework with unsupervised Joint training of CNN (MJ-CNN) for pulmonary DIR. In the proposed multi-scale framework, three CNN models are cascaded, and each works on its own scale level. Each CNN model registers the 3D images by minimizing the dissimilarity between the warped source and the target data and enforcing DVF smoothness, which is technically unsupervised learning. The MJ-CNN network predicts an end-to-end free-form DVF, warping the source data to match the target. To initialize the MJ-CNN, we first trained the three models for their own DIR tasks, sequentially and separately. And then we trained them jointly to achieve an optimal end-to-end registration performance.

To validate MJ-CNN's robustness against the datasets acquired by various scanners and imaging protocols, the MJ-CNN was trained using a public 4D-CT dataset from the SPARE challenge, and then it was tested using the 4D-CT data from the public DIR-LAB dataset, which has landmarks identified on inspiratory and expiratory phases. The registration accuracy of the MJ-CNN was compared to the iterative optimization-based methods including Elastix (Staring *et al* 2010) and a commercial deformable multi-pass B-Spline algorithm provided in Velocity AI (version 3.2.1, Varian Medical Systems, Palo Alto, CA, hereafter referred to as the 'Velocity'), as well as other recently published deep learning-based methods for DIR (Eppenhof and Pluim 2018b, de Vos *et al* 2019). Furthermore, to further validate the generalization of the MJ-CNN across different tasks and imaging techniques, we evaluated the trained MJ-CNN on the clinical CT-to-CBCT and CBCT-to-CBCT registration, and compared it against the Velocity on the same testing data as a baseline comparison. Note that once the MJ-CNN was trained on the SPARE 4D-CTs, no re-training or fine-tuning was performed for all the evaluations mentioned above.

2. Materials and methods

2.1. Problem formulation

The DIR problem can be described as finding a function f to estimate the DVF between the source image data S and the target image data T so that

$$\arg\min_{f} \left[D\left(\phi \cdot S,\ T\right) \ + \ \ast R(\phi)\right] \tag{1}$$

where ϕ is the DVF between S and T, $\phi \cdot S$ is the S deformed by ϕ, D is the dissimilarity between $\phi \cdot S$ and T, R is a regularization item to penalize the local spatial variations in the ϕ. λ is the trade-off parameter for different tasks. And large λ encourages smooth DVFs.

Instead of a computationally-expensive optimization process for each test data, in this study, the function f is defined as a network of cascaded CNN models to estimate DVF at multi-scale levels, and it is optimized during the training process.

2.2. Multi-scale DIR framework with unsupervised Joint training of CNN models (MJ-CNN)

2.2.1. *The Multi-scale Framework of MJ-CNN*

The multi-scale strategy has shown effectiveness in many previous studies for flow estimation (Caballero *et al* 2017, Sokooti *et al* 2017, de Vos *et al* 2019). In this study, multiple CNN models were chained and they took paired data of multiple resolutions as input to estimate DVF under a multi-scale framework, as shown in figure 1.

A coarse-to-fine multi-scale DIR is performed by the proposed MJ-CNN. (1) Scale 1: paired input data S and T are concatenated in the channel dimension and down-sampled by a factor of 4. And then they are fed into the first CNN model for initial DVF estimation. The output of the initial DVF estimation in scale 1 is up-sampled to the original dimension of the input data, and is used to deform the input source data to generate the initially warped source image ($\phi_1 \cdot S$). (2) Scale 2: the CNN model in this scale focuses on registering the residual DVF between the warped source image from scale 1 registration and the target image. Specifically, the warped source image ($\phi_1 \cdot S$) and the original target image (T) are concatenated and down-sampled by a factor of 2, and are then used as the input of the CNN model in scale 2. The residual DVF estimated by scale 2 model is up-sampled to the original input data dimension and is then added to the scale 1 DVF to yield the final scale 2 DVF ($\phi_2 = \phi_1 + \Delta\phi_1$ as shown in figure 1). The scale 2 DVF is used to deform the source image to generate the warped source image ($\phi_2 \cdot S$). (3) Scale 3: the CNN model in this scale focuses on registering the residual DVF between the warped source image from scale 2 and the target image. Specifically, the warped source image ($\phi_2 \cdot S$) and the original target image (T) are concatenated and fed into the CNN model in scale 3. The estimated residual DVF is added to the scale 2 DVF to yield the scale 3 DVF which is the final DVF output of the framework ($\phi_3 = \phi_2 + \Delta\phi_2$ as shown in figure 1).

2.2.2. *Architecture of the CNN model for the DVF estimation*

In this study, a CNN model consisting of convolution, up-sampling, and concatenation layers was proposed to estimate dense DVF between the input paired data. The architecture of the model is shown in figure 2. As indicated by figure 2(A), the model is trained in an unsupervised method, in which no supervised information such as ground truth DVF is required. The model takes 3D paired image data as input and extracts features to estimate the DVF between them. The DVF is then used to deform the source volume to match the target volume using a spatial transformer layer. Thus, the model can update its parameters based on the loss of the dissimilarity between the deformed source and target data under a DVF smoothness constraint.

The CNN model for DVF estimation consists of a contraction path, an expansion path, and the output, as shown in figure 2(B). The contraction path is composed of stacked fully-connected convolution blocks. Convolution with stride 2 is used for the data down-sampling. There are two steps in the expansion path. The first step is a $4 \times 4 \times 4$ up-sampling followed by concatenation with the correspondingly down-sampled features from the contraction path. The second step is a $2 \times 2 \times 2$ up-sampling operation followed by a convolution block, a concatenation with features from the first convolution block in the contraction path, and another convolution block. In each convolution block, a convolutional layer is followed by a leaky rectified linear layer (LeakyRelu) with a leaky rate of 0.2. The output is a convolutional layer without any activation layers. It contains three channels with each representing a 3D vector field component.

Similar to the U-Net (Ronneberger *et al* 2015), a contraction path is used to extract features at multi-scale inception fields with the down-sampling, and in the expansion path, features are up-sampled with concatenating features from the corresponding contraction point to reserve the high-frequency information. However, instead of an expansion path symmetric to the contraction path as adopted in the U-Net, a simple two-step up-sampling path is used in the proposed model to collect features extracted at multi-scale levels from the contraction path,

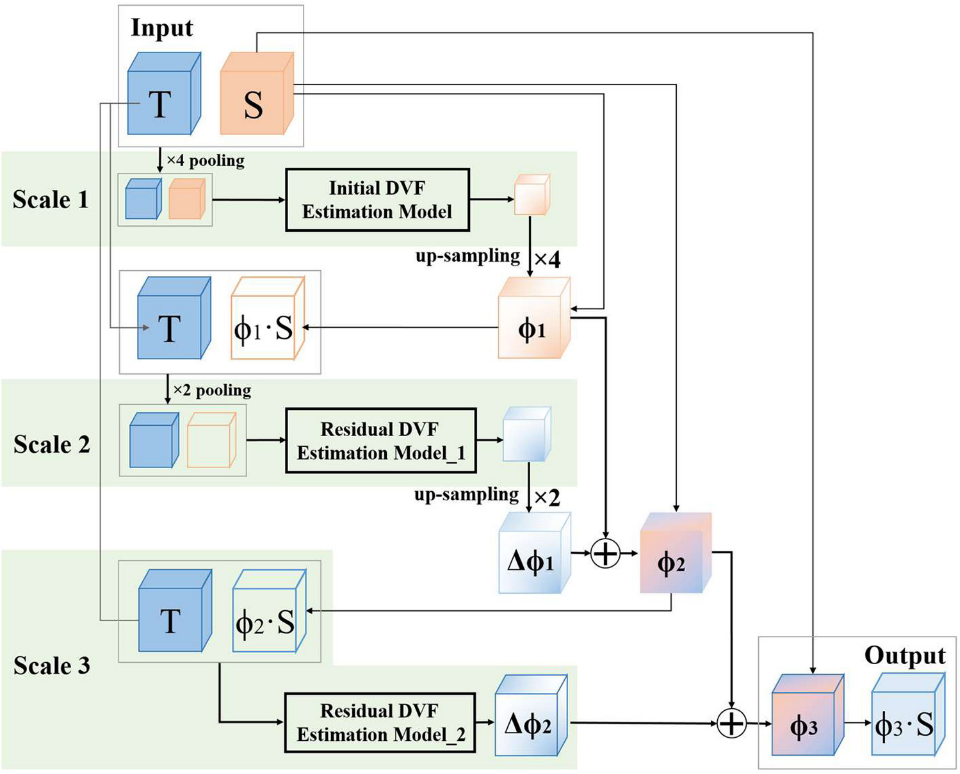

Figure 1. Overall framework of the proposed Multi-scale DIR framework with unsupervised Joint training of CNN models (MJ-CNN), where T indicates the target image, S indicates the source image, ϕ indicates the DVF, $\phi \cdot S$ indicates the S deformed by ϕ, and $\Delta\phi$ indicates the residual DVF between $\phi{\cdot}S$ and T.

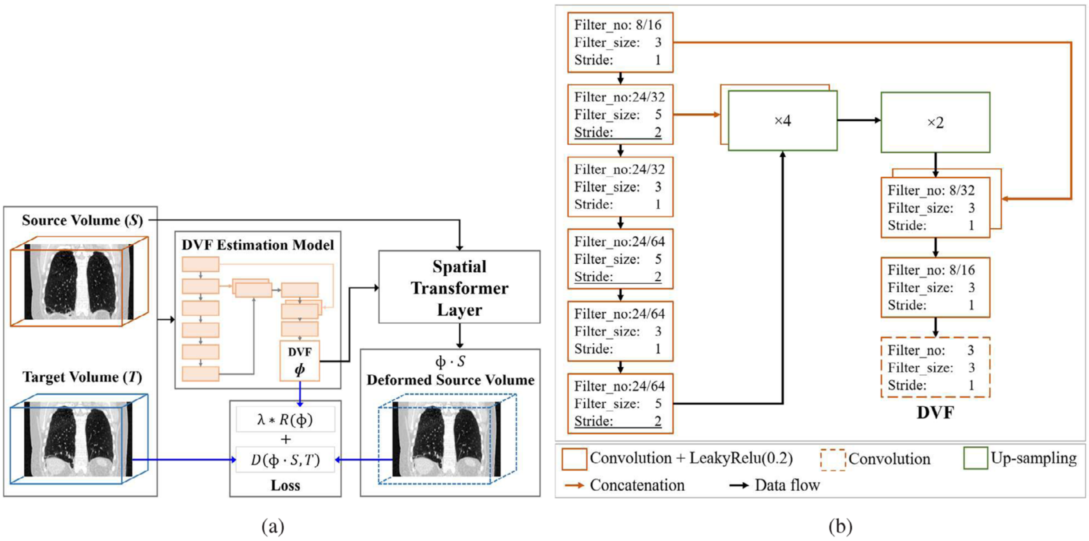

Figure 2. (A) Overall architecture of the proposed CNN model for DVF estimation. R is the regularization on the DVF, and D is the dissimilarity between paired volumes. Details of the DVF estimation model is shown in figure (B). (B) Layers in the proposed CNN model for DVF estimation. In the convolutional layer, filter number is expressed as initial-/residual-DVF estimation model filters. Rectangles represent the output feature maps of the corresponding operations.

and one convolution is performed at the original resolution at last. And then two convolution operations follow to transform features from the image domain to the deformation vector domain to yield the final DVF.

The initial-DVF model in scale 1 and the residual-DVF model in scale 2 and 3 of the MJ-CNN (as shown in figure 1) have the same architecture as presented in figure 2(B). However, compared to the initial-DVF estimation model, the residual-DVF model has more filters in each convolutional layer to deal with the more complicated residual deformation errors.

2.2.3. Joint Training in the MJ-CNN

An important feature of the proposed MJ-CNN is the joint training of the three CNN models at multi-scale levels. Instead of training each CNN model separately, the network trains all three CNN models jointly to minimize the composite loss at multi-scale levels to achieve an overall end-to-end optimal performance.

To initialize the MJ-CNN, models at multi-scale levels were trained, sequentially and separately, to minimize the loss of a weighted sum of a dissimilarity metric and a constraint on DVF smoothness, as shown in equation (2). Variables in equation (2) were defined in equation (1) already.

$$Loss = D\left(\phi \cdot S,\ T\right) \ + \ \ast R(\phi). \tag{2}$$

The scale 1 model was first trained for an initial DVF estimation to match large anatomies. The scale 2 model was then trained to correct residual deformation errors, while fixing the weights of the scale 1 model. The scale 3 model was trained at last to further fine-tune DVF with weights frozen for the scale 1 and 2 models.

After the initialization, all the weights in the MJ-CNN were set to open to update. And a joint training of the individual CNN models in the multi-scale framework was performed in an end-to-end way to synergistically minimize the composite loss at multi-scale levels which is defined as

$$Loss = \ \mu_1 \ast D\left(\phi_1 \cdot S, T\right) + \mu_2 \ast D\left(\phi_2 \cdot S, T\right) + \ \mu_3 \ast D\left(\phi_3 \cdot S, T\right) + \ast R(\phi_3) \tag{3}$$

where S, T, D, R and λ have been defined in equation (1), and ϕ_1, ϕ_2 and ϕ_3 indicate the DVF at scale 1, 2 and 3, respectively, as shown in figure 1, and μ_1, μ_2, μ_3 are weighting factors of the dissimilarity metrics at multi-scale levels.

2.2.4. Configuration of the MJ-CNN

The network proposed in this study learns the deformation mapping between paired image data. During the training process, weights in the network are optimized by minimizing the loss function using Adam (Kingma and Ba 2015) with a learning rate of 10^{-4}.

In equation (2) and (3), the dissimilarity metric D was set to the negative normalized cross correlation, and the DVF regularization item R was set to the l_2-norm DVF gradients (Balakrishnan *et al* 2018). λ was empirically set to 3. For the joint training of the MJ-CNN, μ_1, μ_2 and μ_3 in equation (3) were empirically set to 0.05, 0.05 and 0.9, respectively.

2.3. Experiment design

The proposed MJ-CNN was trained on the public SPARE 4D-CT data. During the sequential training process for initialization, epoch numbers for scale1, 2, and 3 were set to 100, 150, and 150, respectively. During the joint training of the multi-scale framework, 2 cases from the public DIR-LAB dataset were used as validation data to monitor the training process to avoid overfitting.

Performance of the trained MJ-CNN was evaluated both qualitatively and quantitatively on the inter-phase registration of the 4D-CT data from the public DIR-LAB dataset, and on the CT-to-CBCT and CBCT-to-CBCT registration using clinical breath-hold lung patient data. Note that once the MJ-CNN was trained on the SPARE 4D-CTs and validated on 2 DIR-LAB cases, no re-training or validation was performed for the evaluations on various datasets and applications.

2.3.1. Training dataset

In this study, training data included 4D-CT of 22 patients with lung cancers which were randomly chosen from the public AAPM SPARE Challenge dataset. Paired image data of the inspiratory and expiratory phases of the 4D-CT were used as the input. Lung regions were manually cropped for each 4D-CT and were resized to volumes of dimension 256 × 256 × 96. Data intensities were clamped to $[-1000, -200]$ HU and scaled to $[0, 0.2]$ to have the model focus on the pulmonary anatomy. By then, a training dataset containing 44 samples (two samples of inspiration-to-expiration and expiration-to-inspiration for each patient) was constructed.

2.3.2. Evaluation on the DIR-LAB 4D-CT

The DIR-LAB dataset was used to evaluate the registration accuracy of the proposed MJ-CNN. It contains ten thoracic 4D-CT data, and each data has a coordinate list of 300 corresponding anatomical landmarks identified on the inspiratory and expiratory phases. Case 3 and case 8 were selected as the validation data, and the other 8 cases were used for evaluation.

The proposed network took paired data of 4D-CT inspiratory and expiratory phases as input, and predicted the DVF as output. Registration accuracy was evaluated by calculating the landmark registration errors. For quantitative evaluation, registration error (RE) is defined as the l_2-norm of the difference between the deformed landmark and its corresponding reference landmark, as shown in equation (4).

$$\mathrm{RE}\,(l) = \|l_t + \ v - \ l_s\| \tag{4}$$

where l_s and l_t are the corresponding landmarks on the source and target phase, and v is the predicted DVF at the position of the l_t. Jacobian determinant for every point in the DVF was calculated to evaluate the image folding. A negative Jacobian determinant indicated the singularity where the image folding occurred.

To validate the effects of the number of scale levels as well as the base model in each scale, experiments were performed on various network framework settings. First, with 3 scale levels, base model was set to (1) the U-Net model proposed in VoxelMorph (Balakrishnan *et al* 2018) and (2) our proposed model mentioned in section 2.2.2 to evaluate how the base model influenced the network performance. And then, using the base model with superior performance, scale level number was set to 2, 3, and 4 to study the relationship between the scale levels and the network performance. Detailed configurations of the compared network frameworks can be found in appendix A.

To validate the improvements made by the proposed network, results recently reported by two other deep learning methods including the DIR-3D-UNet (Eppenhof and Pluim 2018a) and the DLIR (de Vos *et al* 2019) were included. Registration errors of the conventional iterative optimization-based methods, including Elastix (Staring *et al* 2010) and Velocity, were used as the comparison baseline. Note that, to avoid bias in the comparison, we did not reproduce the results for the DIR-3D-UNet (Eppenhof and Pluim 2018a) and the DLIR (de Vos *et al* 2019), as they may not be consistent with the best results reported in the original papers due to the impact of various factors affecting the model performance. Instead, we cited the reported results of the comparing methods from their original papers for a direct comparison. Registration results of the Elastix included in this study were reported in Eppenhof and Pluim (2018b), in which Elastix performed the registration using the parameters published by Staring *et al* (2010).

2.3.3. Evaluation on the Registration of CT-to-CBCT and CBCT-to-CBCT

To validate the generalization of the features extracted by MJ-CNN across various applications and imaging techniques, registrations between the clinical planning CT and on-board CBCT as well as between on-board CBCTs from different days were performed. In this study, we enrolled 6 lung cancer patients at our institution treated with breath-hold and scanned with 3D planning CT and 3D CBCT. For each patient, data including a planning CT and two CBCTs from different days were collected under an IRB-approved protocol.

Rigid transformation was first performed between the planning CT and on-board CBCT as well as the CBCTs from different days. And then the source and target data were concatenated in the channel dimension and were fed into the proposed network for DVF estimation. Deformed source image data were directly output by the network, and the target image data were used as the ground truth for evaluation. Registration results were evaluated in both the DVF domain and the image domain. Qualitatively, structure alignment and image folding were inspected visually. Quantitatively, structure similarity (SSIM) (Jiang *et al* 2019), peak signal-to-noise ratio (PSNR) (Jiang *et al* 2019) and normalized cross correlation (NCC) (Zhang *et al* 2015) were calculated within the pulmonary regions to evaluate the similarity between the deformed source and the target images, and Jacobian determinant of DVF was calculated to evaluate the degree of image folding.

For the image similarity evaluation, instead of calculating the metrics from the whole sparse lung regions, we extracted patches of dimension $17 \times 17 \times 3$ centered at the feature points on the lung tissues. For each testing case, 1000 feature points were automatically extracted on the target images within the clinical lung contours using a corner feature extraction algorithm (Harris and Stephens 1988) provided by Matlab 2019a. Besides, metrics were calculated for the clinical PTV to evaluate the tumor region alignment.

As the comparison baseline, the Velocity, a commercial iterative optimization-based multi-pass DIR algorithm, was performed on the same testing data.

3. Results

3.1. DIR-LAB landmark errors

3.1.1. Evaluations on network framework settings

Case 3 and case 8 in the DIR-LAB dataset were used for validation, and landmark registration errors of the other eight testing cases were shown in table 1. Results in the base model section (column 3 and 4) showed that, as the base models of the 3-scale network, our proposed model demonstrated superior registration accuracy compared to the U-Net for all the testing cases. Thus, the base model of the MJ-CNN was set to the proposed model in this study. Results in the scale level section (column 5, 6 and 7) showed that registration accuracy improved with the scale level number increased, especially from 2-scale to 3-scale for the cases with large initial deformations over 10 mm. The 3-scale and 4-scale networks showed an overall comparable registration accuracy. Yet the 4-scale network consumed much more computing resources than the 3-scale network. As a result, the MJ-CNN was set to 3 scale levels using the proposed model as the base model. Representative slices are shown in figures 3 and 4.

Table 1. Results of the evaluations on network framework settings using DIR-LAB dataset. Registration errors are expressed as average ± standard deviation in units of mm.

		Base model[a]		Scale levels[b]		
Case	Before registration	U-Net	Proposed	2-scale	3-scale	4-scale
1	3.89 ± 2.78	1.53 ± 0.81	1.20 ± 0.63	1.19 ± 0.64	1.20 ± 0.63	1.15 ± 0.61
2	4.34 ± 3.90	1.42 ± 0.89	1.13 ± 0.56	1.17 ± 0.62	1.13 ± 0.56	1.12 ± 0.60
4	9.83 ± 4.85	2.56 ± 1.66	1.55 ± 0.96	1.63 ± 1.06	1.55 ± 0.96	1.59 ± 1.01
5	7.48 ± 5.50	2.61 ± 1.79	1.72 ± 1.28	1.84 ± 1.30	1.72 ± 1.28	1.71 ± 1.25
6	10.89 ± 6.96	3.79 ± 2.62	2.02 ± 1.70	2.58 ± 2.24	2.02 ± 1.70	1.92 ± 1.59
7	11.03 ± 7.42	3.41 ± 2.62	1.70 ± 1.03	1.92 ± 1.33	1.70 ± 1.03	1.71 ± 1.02
9	7.92 ± 3.97	3.89 ± 2.49	1.51 ± 0.94	1.52 ± 0.82	1.51 ± 0.94	1.53 ± 0.77
10	7.30 ± 6.34	2.54 ± 2.14	1.79 ± 1.61	1.89 ± 1.75	1.79 ± 1.61	1.72 ± 1.47
All	7.83 ± 6.00	2.72 ± 2.18	1.58 ± 1.19	1.72 ± 1.39	1.58 ± 1.19	1.56 ± 1.13

[a] Results of the evaluations on the effects of base model with scale level number of 3.

[b] Results of the evaluations on the effects of scale level numbers with base model of superior performance. In this study, the base model was set to our proposed model in section 2.2.2.

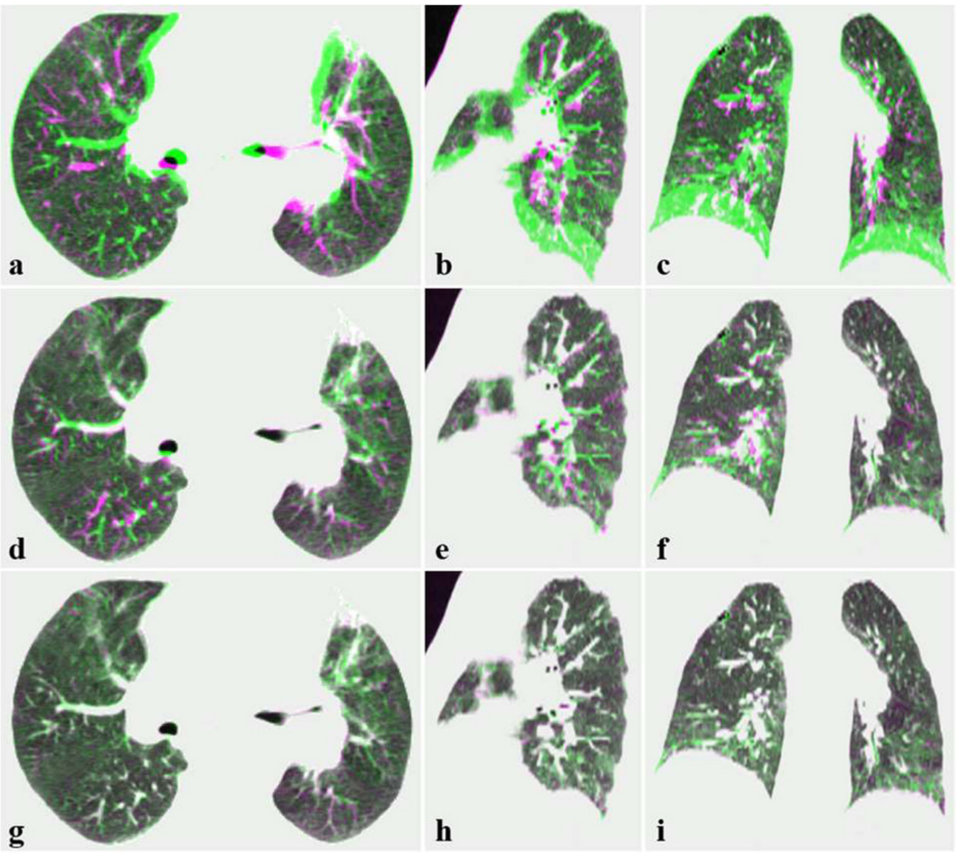

Figure 3. Representative slice of the DIR-LAB Case 6 registration results of (a)–(c) before registration, and the 3-scale MJ-CNN with base model of (d)–(f) U-Net and (g)–(i) the proposed model, with the (deformed) source phase in pink and the target phase in green. Display range is set to [−1000, −200] HU to focus on the lung tissues.

3.1.2. Comparison with other DIR methods

Landmark registration errors of the ten cases in the DIR-LAB dataset were shown in table 2. Results showed that landmark errors were considerably reduced after the DIR for all cases with all the methods listed in the results. Compared to the Elastix-using-mask which is an iterative DIR method aided by the lung mask, MJ-CNN achieved comparable registration accuracy while requiring no lung mask during the evaluation. And compared to the iterative DIR methods without using lung masks, including the Elastix-no-mask and the Velocity, MJ-CNN demonstrated smaller registration errors. Compared to the other two deep learning-based DIR networks of the DIR-3D-UNet and the DLIR, MJ-CNN showed superior DIR accuracy in the average landmark registration

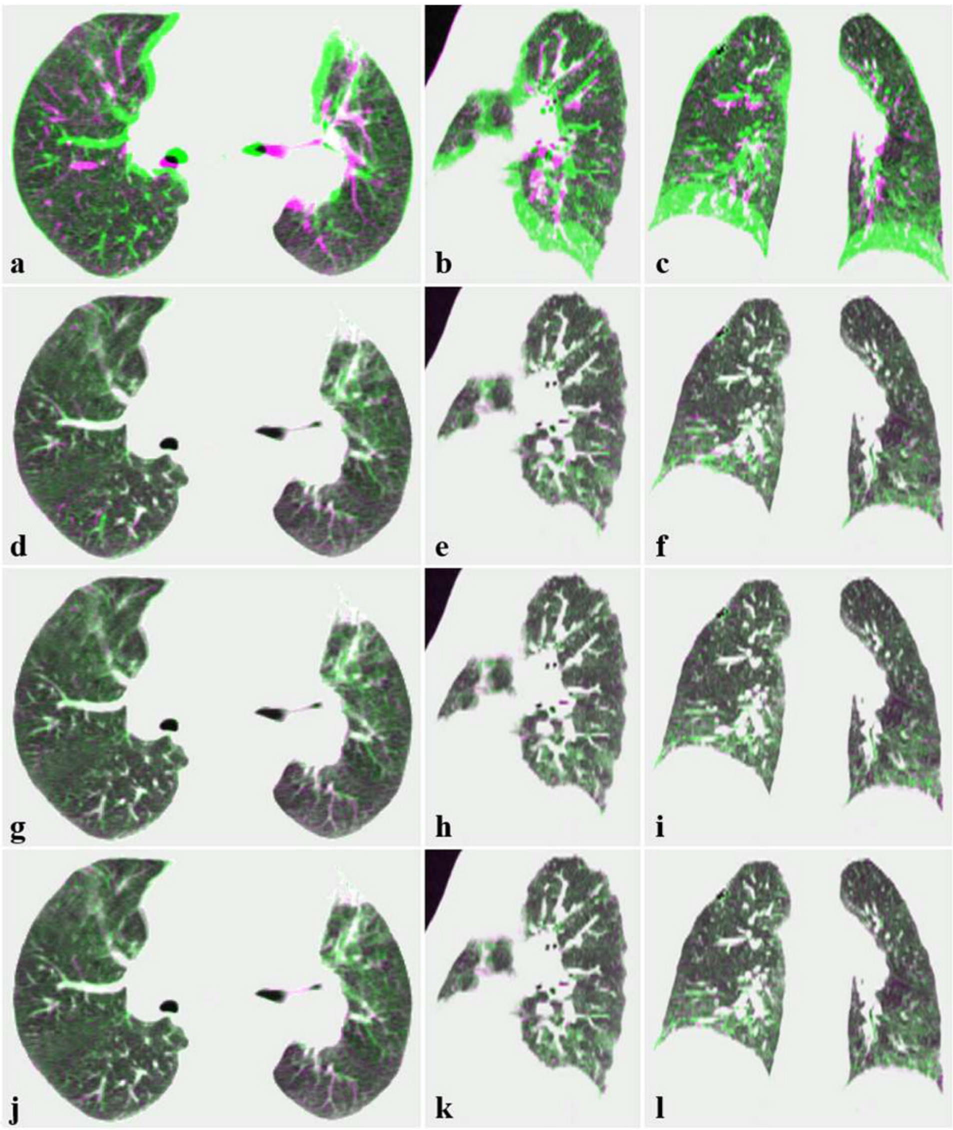

Figure 4. Representative slice of the DIR-LAB Case 6 registration results of (a)–(c) before registration, (d)–(f) 2-scale, (g)–(i) 3-scale and (j)–(l) 4-scale MJ-CNN, with the (deformed) source phase in pink and the target phase in green. Display range is set to [−1000, −200] HU to focus on the lung tissues.

accuracy for all the ten cases, demonstrating the effectiveness of the proposed multi-scale framework with joint training.

For the DIR-3D-UNet, its results were not evaluated on the original data dimension, since the method required resizing the input data to dimensions of 128 × 128 × 128. Even though the study demonstrated that resizing the data had little effect on the registration accuracy on the Elastix (Staring *et al* 2010), no evaluation on the resizing effect for the DIR-3D-UNet was conducted. In contrast, our proposed MJ-CNN can take data of any dimensions as input, and the results in this study were evaluated on the original data resolution. What's more, MJ-CNN's average registration errors of the Case 2, 4 and 9 were even smaller than the reported results of DIR-3D-UNet which was trained on the DIR-LAB dataset itself.

Compared to the DLIR method, the proposed MJ-CNN showed substantial improvements in the registration accuracy especially for the cases with large initial deformations over 10 mm. Note that the DLIR was trained and tested on the DIR-LAB dataset using the leave-one-out strategy, while the proposed MJ-CNN was trained on the SPARE dataset, validated on two cases and tested on the other eight cases of the DIR-LAB dataset. This demonstrated the superior generalization and robustness of the proposed MJ-CNN against scanners and imaging protocols.

3.2. Registration of CT-to-CBCT and CBCT-to-CBCT

Pairs of CT and CBCT as well as CBCTs from different days were pre-aligned using rigid registration and were then fed into the proposed network. Deformed source image data was directly output by the network and was

Table 2. Registration errors evaluated on the DIR-LAB dataset. More results from the published literatures are provided on the DIR-LAB official site (www.dir-lab.com/Results.html). Results are expressed as average ± standard deviation in units of mm.

		Elastix			DIR-3D-UNet			
Case	Before registration	Using mask[a]	No mask[b]	Velocity[c]	CREATIS[d]	DIR-LAB[e]	DLIR[f]	Proposed MJ-CNN
1	3.89 ± 2.78	0.99 ± 0.57	1.04 ± 0.51	1.53 ± 0.64	1.45 ± 1.06	—	1.27 ± 1.16	1.20 ± 0.63
2	4.34 ± 3.90	0.94 ± 0.53	1.20 ± 0.96	1.42 ± 1.13	1.46 ± 0.76	1.24 ± 0.61	1.20 ± 1.12	1.13 ± 0.56
3	6.94 ± 4.05	1.13 ± 0.64	1.76 ± 1.49	1.89 ± 1.23	1.57 ± 1.10	—	1.48 ± 1.26	1.30 ± 0.70[g]
4	9.83 ± 4.85	1.49 ± 1.01	1.73 ± 1.57	2.33 ± 1.23	1.95 ± 1.32	1.70 ± 1.00	2.09 ± 1.93	1.55 ± 0.96
5	7.48 ± 5.50	1.77 ± 1.53	2.42 ± 2.74	2.26 ± 1.73	2.07 ± 1.59	—	1.95 ± 2.10	1.72 ± 1.28
6	10.89 ± 6.96	1.29 ± 0.85	1.98 ± 1.59	2.67 ± 2.20	3.04 ± 2.73	—	5.16 ± 7.09	2.02 ± 1.70
7	11.03 ± 7.42	1.26 ± 1.09	2.90 ± 3.68	3.78 ± 4.39	3.41 ± 2.75	—	3.05 ± 3.01	1.70 ± 1.03
8	14.99 ± 9.00	1.87 ± 2.57	5.10 ± 7.48	5.71 ± 6.12	2.80 ± 2.46	—	6.48 ± 5.37	2.64 ± 2.78[g]
9	7.92 ± 3.97	1.33 ± 0.98	1.81 ± 1.51	2.94 ± 2.03	2.18 ± 1.24	1.61 ± 0.82	2.10 ± 1.66	1.51 ± 0.94
10	7.30 ± 6.34	1.14 ± 0.89	1.79 ± 1.95	2.72 ± 2.64	1.83 ± 1.36	—	2.09 ± 2.24	1.79 ± 1.61
All	8.46 ± 6.58	1.32 ± 1.24	2.17 ± 3.22	2.73 ± 3.07	2.17 ± 1.89	—	2.64 ± 4.32	1.66 ± 1.44

[a] Results of the Elastix using lung mask.

[b] Results of the Elastix without lung mask.

[c] Results of the Velocity AI (version 3.2.1, Varian Medical Systems, Palo Alto, CA) using a deformable multi-pass B-Spline algorithm.

[d] Results of the DIR-3D-UNet trained on the CREATIS dataset.

[e] Results of the DIR-3D-UNet trained on the DIR-LAB dataset.

[f] Results of the DLIR trained on the DIR-LAB dataset using leave-one-out scheme.

[g] Results of the validating cases.

compared to the target. Deformable registrations for the same input were also performed by Velocity using the deformable multi-pass B-Spline algorithm, serving as the comparison baseline.

Figure 5 showed the CT-to-CBCT and CBCT-to-CBCT registration results. Quantitatively, MJ-CNN showed superior performance to the Velocity in the feature point patch and the PTV analysis for both the CT-to-CBCT and the CBCT-to-CBCT registration.

Qualitatively, lung tissues were well matched for all the methods. Figure 6 showed a representative slice of the registration between the planning CT and the on-board CBCT. Figure 7 showed a representative slice of the registration between the on-board CBCTs from different days. All the methods deformed the tumor to the correct position, yet MJ-CNN showed less error within the PTV as well as around the lung vessels.

In this evaluation, to focus on the registration within the pulmonary regions, the outside-lung regions were cropped using the clinical lung mask for quantitative analysis and visual inspection. Quantitative results showed that, compared to the Velocity, an overall improvement in the registration accuracy was made by the proposed MJ-CNN in both structure similarity and HU differences. Qualitative results indicated that, compared to the Velocity, MJ-CNN deformed source data showed less error around lung tissues, and MJ-CNN demonstrated more accurate tumor alignment, which are supported by figure 5.

3.3. Image folding of the deformation

Image folding was detected when the Jacobian determinant of the DVF is negative. For all the testing cases in this study, the proposed MJ-CNN yielded less than 0.1% voxels that showed image folding. Most folding voxels showed up at the lung boundaries where sliding motion can occur. The results indicated the excellent anti-folding performance of the proposed multi-scale network.

3.4. Runtime

The network was implemented in Python (v3.5) with Keras (v2.2.4) framework using TensorFlow (v1.11.0) backend, and experiments were performed on a computer equipped with a GPU of NVIDIA Quadro P4000 and CPU of Intel Xeon with 64GB memory. The prediction including estimating DVF and deforming the source data took around 1.4 s on GPU for paired volumes of dimension 256 × 256 × 96, and the prediction time is linear to the size of the input data. The prediction is nearly real-time and requires no manual-tuning of parameters, making the proposed MJ-CNN very applicable for clinical tasks.

4. Discussion

Intra-patient pulmonary CT registration is an important clinical practice for effective diagnosis and treatment of the lung diseases, which aligns the lung and its internal structures. In this study, we presented a multi-scale DIR

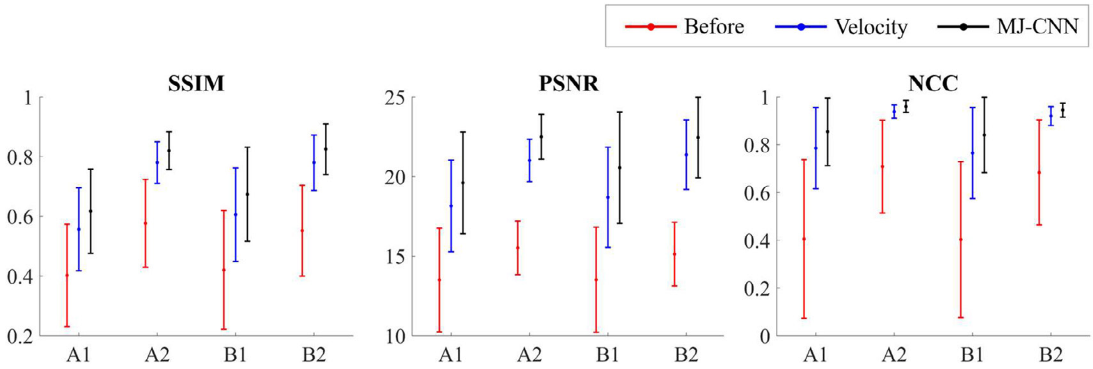

Figure 5. Results of CT-to-CBCT and CBCT-to-CBCT registration in error-bar expressing mean ± standard deviation. (A1) is the results of the CT-to-CBCT feature point patches analysis, (A2) is the results of CT-to-CBCT PTV analysis, (B1) is the results of the CBCT-to-CBCT feature point patches analysis, and (B2) is the results of the CBCT-to-CBCT PTV analysis.

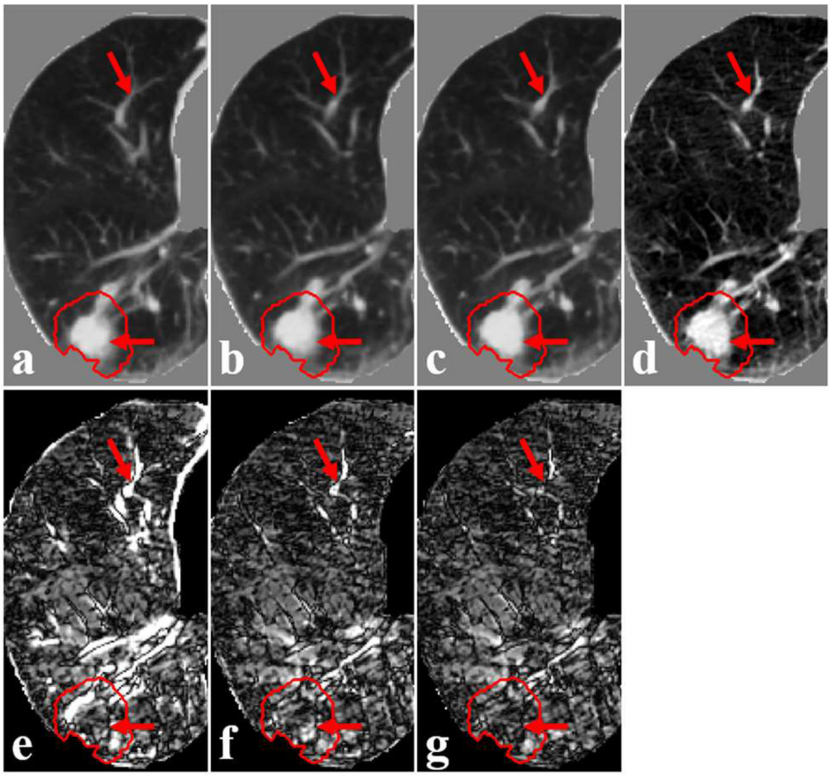

Figure 6. Representative slices of the registration between the planning CT and on-board CBCT. (a) is the planning CT, (b) is the CT deformed by Velocity, (c) is the CT deformed by the proposed network (MJ-CNN), (d) is the target CBCT, and (e)–(g) are the corresponding difference images between (a)–(c) and the target (d). Clinical PTV is highlighted in red solid lines for reference. Red arrows indicate image details for visual inspection. Display range of (a)–(e) is [−1000, 200] HU, and display range of (f)–(i) is [0, 300] HU.

framework with unsupervised joint training of CNN for pulmonary CT registration, which extracts features from the input paired 3D data and directly learns the dense DVF between them in an end-to-end method.

The proposed network showed good accuracy for 4D-CT inter-phase, CT-to-CBCT and CBCT-to-CBCT deformable registration. Compared to the conventional iterative optimization-based methods, it has two major advantages. Firstly, it replaces the computationally-expensive and time-consuming optimizing process for the individual testing case with a network optimization over a dataset during the training process. Instead of iteratively searching for an optimal deformation pattern for each test data, the proposed method directly applies the learned pattern to predict the DVF for the test data with nearly real-time speed. Secondly, to achieve accurate DIR accuracy, manual-tuning of parameters is often required for individual testing case in some conventional iterative optimization-based methods (Avants *et al* 2009, Modat *et al* 2010), making the methods user dependent and time-consuming. For the proposed MJ-CNN, the parameterization is optimized during the training process.

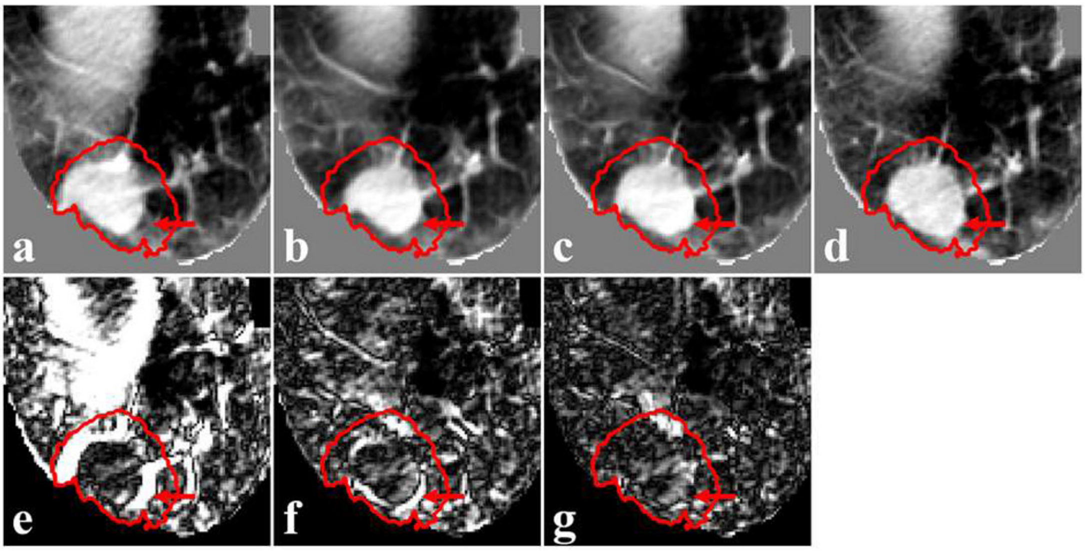

Figure 7. Representative slices of the registration between the on-board CBCT from different days. (a) is the source CBCT, (b) is the source deformed by velocity, (c) is the source deformed by the proposed network (MJ-CNN), (d) is the target CBCT, and (e)–(g) are the corresponding difference images between (a)–(c) and the target (d). Clinical PTV is highlighted in red solid lines for reference. Display range of (a)–(d) is [−1000, 200] HU, and display range of (e)-(g) is [0, 300] HU.

And once the network is trained, no manual-tuning of parameters is required for each testing case, making the registration process fully automatic, user-independent and fast.

The proposed MJ-CNN was tested using the public DIR-LAB dataset, and its registration accuracy was quantitively evaluated on the registration landmark errors (table 2). Compared to the conventional iterative optimization-based methods, the proposed MJ-CNN substantially reduced the registration time to a few seconds (section 3.4) with comparable or even superior DIR accuracy. Compared to the two recently published deep learning-based DIR methods of the DIR-3D-UNet (Eppenhof and Pluim 2018b) and DLIR (de Vos *et al* 2019), the proposed MJ-CNN achieved the smallest registration errors. Such preferable performance of MJ-CNN may result from the following facts.

For comparison with the DIR-3D-UNet, the DIR-3D-UNet is a supervised learning-based method, and therefore its accuracy is affected by the accuracy of the ground truth DVF synthesized for training. And the model was trained to minimize the difference between the predicted DVF and the ground truth DVF without a regularization constraint. In comparison, the proposed MJ-CNN was directly optimized based on the endpoint of minimizing the dissimilarity between the warped source and the target images, based on unsupervised learning. And a regularization constraint on the DVF smoothness was used during the training process of the MJ-CNN, making the predicted DVF more realistic. Besides, the DIR-3D-UNet performed the DIR on a single scale, which is more prone to local minimums for the pulmonary regions with complicated and large deformations. Such problem can be alleviated by the multi-scale framework adopted by the MJ-CNN.

For comparison with the DLIR, although DLIR also adopted the multi-scale framework, its registration accuracy for the pulmonary cases with large deformations is limited, as shown in table 2. Potential reasons are in two aspects. (1). The DLIR method predicts the DVF at control points using a patch-based approach, where paired image patches are extracted from the source and target images and are used to estimate the control point deformation vectors located at the center of patches. In such patch-based methods, the patch size needs to be optimized based on the deformation magnitude. And the performance of the DIR models can be degraded when the patches are too small compared to the deformation magnitude, leading to insufficient context for the registration, or too large with much distant information, leading to an inaccurate representation of the local deformation at the center point (Eppenhof and Pluim 2018b). However, the deformation magnitude is unknown before the registration and varies within individual cases as well as over different cases. As a result, predefined patch sizes are often not optimal, leading to errors in the registration. Besides, the control points in the DLIR method are located uniformly throughout the images, which may not sample the DVF finely enough at regions with a high gradient of deformation, leading to errors in the dense DVF generation afterward. In contrast, the proposed MJ-CNN adopted CNN models for free-form dense DVF estimation based on the entire volumes. Image information available for estimating deformation vector at each voxel is no longer limited to a predefined patch range. As shown in table 2, MJ-CNN is robust in registering both small and large deformations. (2). In the training process of the DLIR, the CNN model at each resolution level was trained sequentially and separately. Although each model may be trained to achieve the optimal performance at the corresponding resolution level, their overall performance when combined in the framework may not be optimized to achieve the accurate end-to-end registration results. In the proposed MJ-CNN, a joint training was performed to train the CNN models at multi-scale levels together to improve the end-to-end registration accuracy.

The generalization of the features extracted by the proposed MJ-CNN was evaluated on the registration between the planning CT and on-board CBCT as well as on-board CBCTs from different days. Note that, during this evaluation, no re-training or fine-tuning was conducted for the MJ-CNN which was trained on the 4D-CT. Since lung tissues are relatively sparse, the metrics calculated from the entire lung regions are not sensitive to the DIR errors. To address this, small patches centered at the lung tissue feature points were extracted for the evaluation. MJ-CNN trained on the 4D-CT dataset performed well on the registration between data scanned on different days using different imaging techniques with various resolutions and noise levels. Although only six patients with 12 image pairs were included in this study due to the limited clinical data access, the results preliminarily demonstrated the generalizing ability of the proposed MJ-CNN across various applications and imaging techniques, indicating that MJ-CNN was able to extract the underlying features for the deformation prediction.

The proposed MJ-CNN was optimized on a composite loss composed of similarity in the image domain as well as the smoothness in the DVF domain. However, sliding motion around the lung boundary yielded folding in the DVF, which competed against the DVF smoothness regularization term R in equations (2) and (3). In future studies, the sliding effect around the lung boundary can be specifically modeled to further improve the DIR performance. For example, adaptive weights of the DVF bending energy penalty around the lung boundaries can be incorporated in the loss function.

5. Conclusion

The proposed multi-scale DIR framework with unsupervised joint training of CNN is effective and efficient in 4D-CT, CT-to-CBCT and CBCT-to-CBCT deformable registration, and requires no manual-tuning of parameters during prediction. It can become a very valuable tool for various clinical tasks.

Acknowledgment

This work was supported by the National Institutes of Health under Grant No. R01-CA184173 and R01-EB028324. The authors acknowledge the financial support from the Chinese Scholarship Council (CSC).

Appendix

Figure 1 of the manuscript shows an overall framework of the proposed Multi-scale DIR framework with unsupervised Joint training of CNN models (MJ-CNN). All the compared networks were trained on the public SPARE dataset, validated on case 3 and case 8 of the DIR-LAB dataset, and tested on the other 8 cases of the DIR-LAB dataset. Detailed configurations are as follows.

A.1. Evaluation on the base model of the MJ-CNN

For the 3-scale MJ-CNN(U-Net), the base model for every single scale was set to the U-Net proposed in the VoxelMorph (Balakrishnan *et al* 2018). Other configurations were the same as mentioned in Section II.B of the manuscript. For the training, sequential training for initialization took 100, 150 and 150 epochs for scale 1, 2 and 3, respectively, and the joint training process took 150 epochs and was monitored by the validation data.

For the 3-scale MJ-CNN(proposed), the configurations were in the section 2.2 of the manuscript. For the training, sequential training for initialization took 100, 150 and 150 epochs for scale 1, 2 and 3, respectively, and the joint training process took 150 epochs and was monitored by the validation data.

A.2. Evaluation on the scale level number of the MJ-CNN

For the 2-scale MJ-CNN, base model for scale 1 was set to the initial-DVF estimation model, and base model for scale 2 was set to the residual-DVF estimation model, shown in figure 2(B) of the manuscript. Down-sampling factors were set to 2 and 1 for scale 1 and 2, respectively. In the loss function, μ_1 and μ_2, the weighting factors of the dissimilarity metrics at multi-scale levels were set to 0.1 and 0.9, respectively. Other configurations were the same as mentioned in section 2.2 of the manuscript. For the training, sequential training for initialization took 100 and 150 epochs for scale 1 and 2, respectively, and the joint training process took 150 epochs and was monitored by the validation data.

Configurations of the 3-scale MJ-CNN can be found in the sections 2.2 and 2.3 of the manuscript.

For the 4-scale MJ-CNN, base model for scale 1 was set to the initial-DVF estimation model, and base model for scale 2, 3 and 4 was set to the residual-DVF estimation model, shown in figure 2(B) of the manuscript. Down-sampling factors were set to 8, 4, 2 and 1 for scale 1, 2, 3 and 4, respectively. In the loss function, μ_1, μ_2, μ_3 and μ_4, the weighting factors of the dissimilarity metrics at multi-scale levels were set to 0.03, 0.03, 0.04 and 0.9, respectively. Other configurations were the same as mentioned in section 2.2 of the manuscript. For the training, sequential training for initialization took 100, 150, 150 and 150 epochs for scale 1, 2, 3 and 4, respectively, and the joint training process took 150 epochs and was monitored by the validation data.

References

Avants B B, Tustison N and Song G 2009 Advanced normalization tools (ANTS) *Insight J.* **2** 1–35

Balakrishnan G, Zhao A, Sabuncu M R, Guttag J and Dalca A V 2018 An unsupervised learning model for deformable medical image registration *Proc. of the IEEE Conf. Comput. Vis. Pattern Recognit.* pp 9252–60

Caballero J, Ledig C, Aitken A, Acosta A, Totz J, Wang Z and Shi W 2017 Real-time video super-resolution with spatio-temporal networks and motion compensation *Proc. of the IEEE Conf. Comput. Vis. Pattern Recognit.* pp 4778–87

Cao X, Yang J, Zhang J, Nie D, Kim M, Wang Q and Shen D 2017 Deformable image registration based on similarity-steered CNN regression *Int. Conf. Med. Image Comput. Comput. Assisted Intervention* (Berlin: Springer) pp 300–8

de Vos B D, Berendsen F F, Viergever M A, Sokooti H, Staring M and Išgum I 2019 A deep learning framework for unsupervised affine and deformable image registration *Med. Image Anal.* **52** 128–43

Eppenhof K A and Pluim J P 2017 Supervised local error estimation for nonlinear image registration using convolutional neural networks *Proc. SPIE* **10133** 101331U

Eppenhof K A and Pluim J P 2018a Error estimation of deformable image registration of pulmonary CT scans using convolutional neural networks *J. Med. Imaging* **5** 024003

Eppenhof K A and Pluim J P 2018b Pulmonary CT registration through supervised learning with convolutional neural networks *IEEE Trans. Med. Imaging* **38** 1097–105

Harris C G and Stephens M 1988 A combined corner and edge detector *Alvey Vision Conf.* vol 15 pp 10–5244

Jiang Z, Chen Y, Zhang Y, Ge Y, Yin F-F and Ren L 2019 Augmentation of CBCT reconstructed from under-sampled projections using deep learning *IEEE Trans. Med. Imaging* **38** 2705–15

Kearney V, Haaf S, Sudhyadhom A, Valdes G and Solberg T D 2018 An unsupervised convolutional neural network-based algorithm for deformable image registration *Phys. Med. Biol.* **63** 185017

Kingma D and Ba J 2015 Adam: a method for stochastic optimization (arXiv:1412.6980)

Modat M, McClelland J and Ourselin S 2010 Lung registration using the NiftyReg package. Medical Image Analysis for the Clinic-a Grand Challenge *Proc. 13th Int. Conf. Med. Image Comput. Comput Assist. Intervention (Beijing, China, 20–24 September 2010)* pp 33–42

Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *Int. Conf. Med. Image Comput. Comput. Assist. Intervention* pp 234–41

Shan S, Yan W, Guo X, Chang E I, Fan Y and Xu Y 2017 Unsupervised end-to-end learning for deformable medical image registration (arXiv:1711.08608)

Sokooti H, de Vos B, Berendsen F, Lelieveldt B P, Išgum I and Staring M 2017 Nonrigid image registration using multi-scale 3D convolutional neural networks *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Berlin: Springer) pp 232–9

Staring M, Klein S, Reiber J H, Niessen W J and Stoel B C 2010 Pulmonary image registration with elastix using a standard intensity-based algorithm Medical Image Analysis for the Clinic-a Grand Challenge *Proc. 13th Int. Conf. Med. Image Comput. Comput Assist. Intervention (Beijing, China, 20–24 September 2010)* pp 73–9

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

Yang X, Kwitt R, Styner M and Niethammer M 2017 Quicksilver: fast predictive image registration-a deep learning approach *NeuroImage* **158** 378–96

Zhang Y, Yin F-F, Pan T, Vergalasova I and Ren L 2015 Preliminary clinical evaluation of a 4D-CBCT estimation technique using prior information and limited-angle projections *Radiother. Oncol.* **115** 22–9