Supervised learning in medical image registration

Sokooti, H.

2

Nonrigid Image Registration using Multi-Scale 3D Convolutional Neural Networks

This chapter was adapted from:

H Sokooti, B de Vos, F Berendsen, BP Lelieveldt, I Išgum, and M Staring. **Nonrigid image registration using multi-scale 3D convolutional neural networks**, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017 September.

Abstract

In this chapter, we propose a method to solve nonrigid image registration through a learning approach, instead of via iterative optimization of a predefined dissimilarity metric. We design a Convolutional Neural Network (CNN) architecture that, in contrast to all other work, directly estimates the displacement vector field (DVF) from a pair of input images. The proposed RegNet is trained using a large set of artificially generated DVFs, does not explicitly define a dissimilarity metric, and integrates image content at multiple scales to equip the network with contextual information. At testing time nonrigid registration is performed in a single shot, in contrast to current iterative methods. We tested RegNet on 3D chest CT follow-up data. The results show that the accuracy of RegNet is on par with a conventional B-spline registration, for anatomy within the capture range. Training RegNet with artificially generated DVFs is therefore a promising approach for obtaining good results on real clinical data, thereby greatly simplifying the training problem. Deformable image registration can therefore be successfully casted as a learning problem.

2.1 Introduction

Deformable image registration (DIR) is the task of finding the spatial relationship between two or more images, and is abundantly used in medical image analysis. Typically, image registration is solved by iteratively optimizing a predefined hand-crafted intensity-based dissimilarity metric over the transformation parameters. The metric represents a model of the intensities encountered in the image data. Problems may occur when part of the data does not fit the model, which are typically dealt with by making modifications to the dissimilarity metric. Instead, in this paper we take another approach, where we do not handcraft such a model, but use a machine learning approach to automatically determine what constitutes an accurate registration, i.e. without explicitly defining a dissimilarity metric. The proposed method is based on regression using Convolutional Neural Networks (CNNs), that directly learns a displacement vector field (DVF) from a pair of input images.

The idea of learning registration has shown to be promising [41]. Several CNN regression techniques have been proposed in the context of image registration. Miao *et al.* [42] applied CNN regression for rigid 2D-3D registration. Liao *et al.* [16] used CNN regression to model a sequence of motion actions for 3D registration. Their method is iterative (not one shot), and limited to rigid-body transformations. For nonrigid approaches, Yang *et al.* [43] predicted the initial momentum of a 3D LDDMM registration. Eppenhof *et al.* [44] trained a CNN to predict the local registration error, without performing a full registration. Related work has been done in the field of optical flow [45].

In contrast, we propose an end-to-end method that directly predicts the 3D nonrigid DVF given a fixed and a moving image, without requiring a dissimilarity metric like conventional methods. The proposed architecture, called RegNet, analyzes 3D input patches at multiple scales to equip the CNN with contextual information. Training is based on a wide variety of artificial displacements acting as the target value in the loss function, while testing is performed on registration of baseline and follow-up CT images of a patient. At testing time the registration is performed in a single shot, in contrast to current iterative methods. To the best of our knowledge this is the first method that solves nonrigid 3D image registration with CNNs end-to-end, i.e. directly predicting DVFs.

2.2 Methods

2.2.1 Network architecture

The proposed CNN architecture RegNet takes patches from a pair of 3D images (the fixed image I_F and the moving image I_M) as input. The output of RegNet is a vector of three elements, which is the displacement of the central voxel of the patch. A full DVF

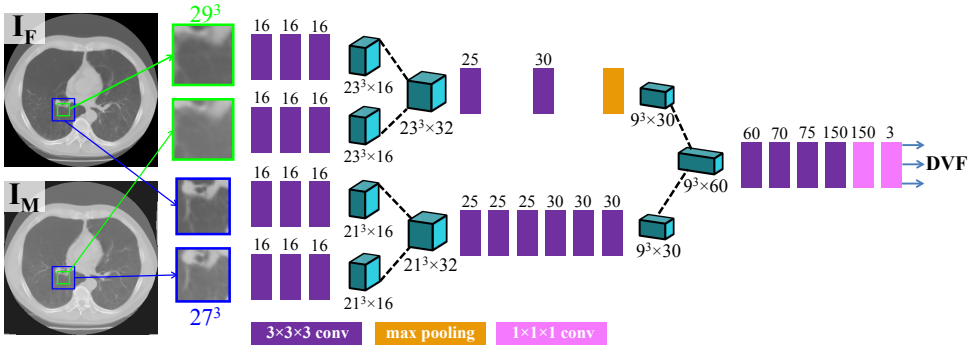


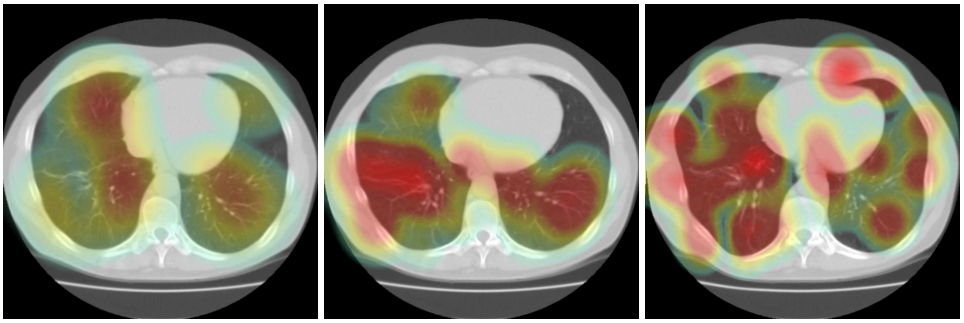
Figure 2.1: RegNet design.

is generated by sliding over the input images. The DVF is defined as the displacement $\boldsymbol{u}(\boldsymbol{x})$, mapping points from the fixed image domain to that of the moving image. The transformation is defined as $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})$.

For each image we extract patches at original resolution of size 29x29x29 voxels. To improve the receptive field of the network, we additionally extract patches of 54x54x54 voxels, which are downsampled to 27x27x27 voxels. In this way local as well as more global information is incorporated, allowing better discrimination between anatomical locations and to add contextual information. The downsampling makes sure there is limited effect on memory consumption and computational overhead. Similar multi-scale approaches have been shown effective for segmentation [46]. We thus have four 3D patches as inputs.

We start with three convolutional layers for each input patch separately (late-fusion) instead of stacking them as channels (early-fusion). The fixed and moving patches of each resolution are then merged by concatenation. This is followed by 2 and 6 convolutional layers for the original resolution and the downsampled patch, respectively. Max pooling is used on the pipeline of the original resolution, ensuring spatial correspondence of the activation of the two pipelines before merging; for every 2 shift of the receptive field of the original resolution only 1 shift should be performed in the low resolution [46]. The two resolution pipelines are then also concatenated, followed by 4 convolutional layers and two fully connected layers. All convolutional layers use $3 \times 3 \times 3$ kernels, batch normalization and ReLu activation. The network architecture is visualized in Fig. 2.1.

Optimization is done using Adam, with a decaying learning rate starting at 0.001 and a decay factor of 1.25 in each epoch, which improved the convergence rate in our experiments. The loss function is defined as the mean residual distance between target and estimated DVF: $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\text{DVF}'_i - \text{DVF}_i|$, with DVF$'$ the prediction of RegNet



(a) Low frequency (b) Medium frequency (c) High frequency

Figure 2.2: Heat maps of the magnitude of DVFs used for training RegNet.

and DVF the target defined in Section 2.2.2.

2.2.2 Training

To train our network, synthetic DVFs are generated with varying spatial frequency and amplitude, aiming to represent the range of displacements that can be seen in real images: 1) Creating a vector field with the size of the input image (which will act as the moving image) and initializing it with zero vectors; 2) Randomly selecting P points in the DVF and randomly assigning three values to the displacement vector in the range $[-\theta, +\theta]$; 3) Smoothing the DVF with a Gaussian kernel with a standard deviation of σ. Low, medium and high frequency deformations are generated using the settings $\sigma = 35, P = 80, \theta = 8$; $\sigma = 25, P = 100, \theta = 8$; and $\sigma = 20, P = 100, \theta = 8$, respectively. Transformed images are generated by applying the DVF to the input image, using cubic B-spline interpolation, resulting in the fixed image. To allow more accurate simulation of real images, Gaussian noise with a standard deviation of 5 is finally added to the images. Examples are available in Fig. 2.2.

It is possible to generate plenty of deformations for a single moving image, but a drawback of this approach is that the moving image is identical in each pair of input images, as only the fixed image is generated randomly. We therefore also generate deformed versions of the moving image, based on which new deformed images are created. The new moving images are generated using low frequency deformations only, to avoid over-stretching (leading to a blurred appearance). We use the settings $\sigma = 35, P = 100, \theta = 8$ and Gaussian noise with a standard deviation of 3 in this step.

2.3 Experiments and results

2.3.1 Materials

We use data from the SPREAD study [47], which contains 19 pairs of 3D chest CT images. The dimension of the images is about $446 \times 315 \times 129$ with an average voxel

size of $0.781 \times 0.781 \times 2.5$ mm. Patients are between 49 and 78 years old and for each patient a baseline image and a 30 months follow-up image are available. For each pair, 100 well-distributed corresponding landmarks were previously selected semi-automatically at distinctive locations [48]. All images were resampled to a voxel size of $1 \times 1 \times 1$ mm.

RegNet is written in Theano [49] and Lasagne [50], artificial DVFs are created using SimpleITK [51]. Conventional registrations are performed using `elastix` [52].

2.3.2 Experimental setup and evaluation

The set of 19 image pairs is divided in a training set of 10 pairs, a validation set of 2 pairs, and a test set of 7 pairs. 2100 patches per image are randomly extracted from the lung regions of the training images, using both the baseline and follow-up images as input for training. For each image in the database we create 6 different DVFs (3 for a single moving image and 3 other after deforming that moving image, see Section 2.2.2), resulting in $252,000$ training examples. In addition, we applied data augmentation, flipping all patches in the x, y and z direction and adding Gaussian noise with a standard deviation of 5. In total we have approximately 1 million patches available for training. The network is trained for 15 epochs. The validation set was used to monitor overfitting during training, and to compare with the single-scale and the early-fusion design.

The test set was used in two ways. We first evaluate the ability of the trained network to register artificially deformed image pairs, which is how RegNet was trained. This was evaluated using the MAE measure. Second, we apply RegNet for registration of the real baseline and follow-up CT images, without artificial deformations. This experiment is evaluated using the set of corresponding landmarks, where we report their mean Euclidean distance after registration: $\text{TRE} = \frac{1}{n}\sum_{i=1}^{n} \|\text{DVF}_i'(\boldsymbol{x}_F) + \boldsymbol{x}_F - \boldsymbol{x}_M\|_2$, with $\boldsymbol{x}_F$ and $\boldsymbol{x}_M$ the landmark locations. An initial affine registration is performed before applying RegNet, similar to conventional approaches. We use an intensity-based method (normalized correlation), using 5 resolutions of 1000 iterations each. RegNet is compared with two conventional B-spline registrations with a final grid spacing of 10mm: a version using a single resolution of 2000 iterations, and one using 3 resolutions of 500 iterations each. As the capture range of our network is certainly less than half the patch width, we additionally present the TRE of only those points that are within 8mm distance after the affine registration (TRE′).

2.3.3 Results

All quantitative results are given in Table 2.1. The results on the validation set show that multi-scale late-fusion RegNet performs better than either single-scale or early-fusion RegNet. It can be seen that the regression accuracy on the validation set (MAE) is about 1 mm, showing that RegNet was successfully trained. The results in the x and

Table 2.1: Quantitative results

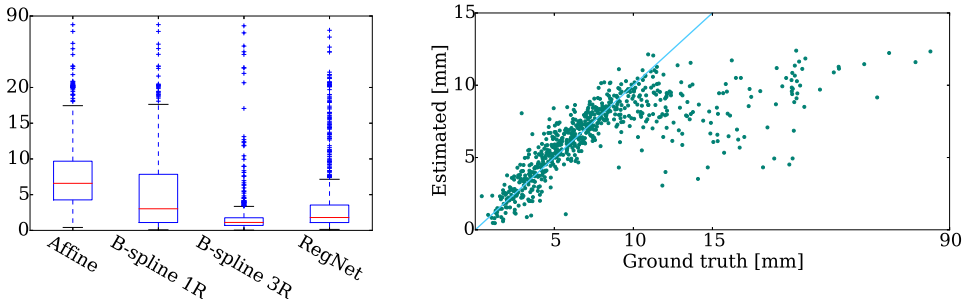
Evaluation	Method	Data	measure	measure$_x$	measure$_y$	measure$_z$
MAE	RegNet 1Scale	validation	1.70 ± 1.81	0.56 ± 0.78	0.53 ± 0.71	0.61 ± 0.88
	RegNet Early	validation	1.26 ± 1.22	0.41 ± 0.51	0.39 ± 0.48	0.45 ± 0.60
	RegNet	validation	1.17 ± 1.10	0.36 ± 0.56	0.38 ± 0.44	0.43 ± 0.49
	RegNet	test	1.19 ± 1.17	0.36 ± 0.59	0.40 ± 0.50	0.43 ± 0.51
TRE	Affine	test	8.08 ± 7.18	4.21 ± 4.40	3.92 ± 5.64	3.80 ± 4.25
	B-spline 1R	test	5.48 ± 7.56	2.47 ± 4.01	2.64 ± 5.71	2.92 ± 4.12
	B-spline 3R	test	2.19 ± 6.22	0.67 ± 1.97	1.04 ± 5.07	1.45 ± 3.21
	RegNet	test	4.39 ± 7.54	2.19 ± 4.53	1.79 ± 4.83	2.35 ± 4.33
TRE′	Affine	test	5.39 ± 2.25	2.80 ± 2.04	2.70 ± 1.92	2.73 ± 1.93
	B-spline 1R	test	2.59 ± 2.28	1.02 ± 1.44	1.09 ± 1.47	1.72 ± 1.56
	B-spline 3R	test	1.28 ± 0.94	0.41 ± 0.51	0.42 ± 0.43	1.00 ± 0.86
	RegNet	test	1.66 ± 1.26	0.58 ± 0.62	0.64 ± 0.77	1.19 ± 1.10

y direction are slightly better than that in the z direction, which can be attributed to the relatively large slice thickness of our data. The MAE results on the test set confirm that RegNet can successfully register artificially deformed images with a sub-voxel accuracy.

For the test set we have 685 corresponding landmarks available to compute the TRE. For TRE′, 503 landmarks are within 8 mm after affine registration. The results for affine, the two B-spline settings and RegNet are listed in Table 2.1 and illustrated in Figs. 2.3 and 2.4. It can be seen that the multi-resolution B-spline method overall gives the best performance (TRE results), but RegNet is better than a single resolution B-spline. When we focus on the points within the capture range of RegNet (TRE′ results) it can be seen that RegNet performs better than the single resolution B-spline method, and performs similar to multi-resolution B-spline. For those landmarks a residual error of 1.7 mm is obtained, which is sub-voxel with respect to the original resolution. Again, the accuracy in the x and y direction is slightly better than that in the z direction. Fig. 2.3b shows a scatter plot of all landmarks after registration with RegNet. RegNet gives accurate registrations until ~8 mm, which is to be expected due to the patch size and the fact that RegNet was trained up to $\theta = 8$ mm deformations only. Figs. 2.4b-d show scatter plots of the landmarks within 8 mm, for the three directions separately. Example registration results are given in Fig. 2.5. Inference time for an image of size 300^3 is about 14 seconds on a Tesla K40.

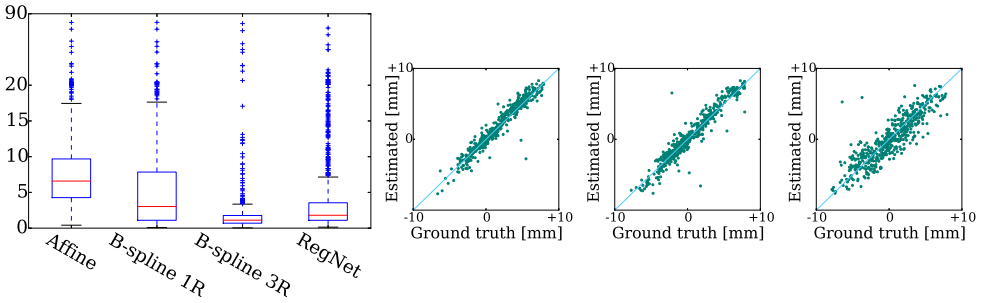
2.4 Discussion and conclusion

We presented a convolutional neural network (RegNet) for 3D nonrigid image registration. RegNet can be successfully applied to real world data, after training on artificially generated displacement vector fields. Tests on artificially deformed images



(a) box-plot (b) Scatter plot for RegNet

Figure 2.3: Residual landmark distances, for all landmarks.

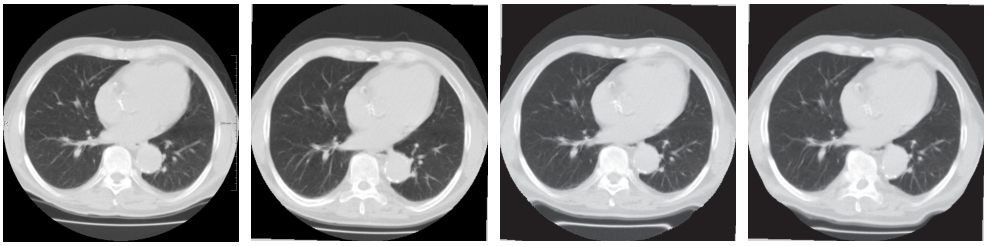


(a) box-plot (b) x direction (c) y direction (d) z direction

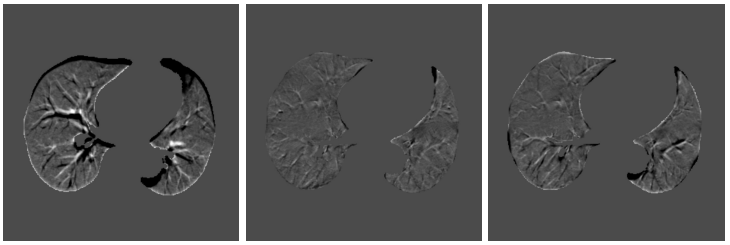
Figure 2.4: Residual landmark distances, for the landmarks in the capture range. Figures b-d show scatter plots of RegNet against the ground truth.

as well as with intra-patient chest CT data, showed that RegNet achieved sub-voxel registration performance, for landmarks within the capture range. This was better than the performance of a conventional single resolution B-spline registration method, and close to that of a multi-resolution B-spline. When considering all landmarks, the multi-resolution B-spline method still outperformed RegNet. In the training phase of RegNet no use was made of (manually annotated) corresponding points, or segmentations for guidance, which are hard to obtain in large quantities. Synthetic DVFs on the other hand can easily be generated in bulk, which greatly simplifies the training process.

In our current design the registration capture range is related to the size of the patches that are shown to the network, and the results show good performance until 8 mm, but deteriorate after that. The capture range may be enlarged by the use of larger patches or the addition of more scales to the network. It is also possible to extend RegNet to a multi-resolution approach, working from even further downsampled (and



(a) Fixed (b) Affine (c) B-spline 3R (d) RegNet



(f) Affine (g) B-spline 3R (h) RegNet

Figure 2.5: Example results (top row) and difference images (bottom row).

smoothed) images than in the current multi-scale approach, successively upsampling until the original resolution.

For future work, we will perform a sensitivity analysis of a number of important parameters of RegNet, like the patch size and its relation to the several parameters that define the training DVFs (e.g. the maximum magnitude θ). We will also train RegNet in other applications besides chest CT, to test the generalizability of the architecture.

In conclusion, the proposed neural network achieves promising results for the nonrigid registration of image pairs, using an end-to-end approach. Information at multiple scales is integrated in the CNN. After training, deformable registration is performed in one shot.

Acknowledgments

This work is financed by the Netherlands Organization for Scientific Research (NWO), project 13351. Dr. M.E. Bakker and J. Stolk are acknowledged for providing a ground truth for the SPREAD study data used in this paper. The Tesla K40 used for this research was donated by the NVIDIA Corporation.