End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network

Bob D. de Vos[1(✉)], Floris F. Berendsen[2], Max A. Viergever[1], Marius Staring[2], and Ivana Išgum[1]

[1] Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands
B.D.deVos-2@umcutrecht.nl

[2] Division of Image Processing, Leiden University Medical Center, Leiden, The Netherlands

Abstract. In this work we propose a deep learning network for deformable image registration (DIRNet). The DIRNet consists of a convolutional neural network (ConvNet) regressor, a spatial transformer, and a resampler. The ConvNet analyzes a pair of fixed and moving images and outputs parameters for the spatial transformer, which generates the displacement vector field that enables the resampler to warp the moving image to the fixed image. The DIRNet is trained end-to-end by unsupervised optimization of a similarity metric between input image pairs. A trained DIRNet can be applied to perform registration on unseen image pairs in one pass, thus non-iteratively. Evaluation was performed with registration of images of handwritten digits (MNIST) and cardiac cine MR scans (Sunnybrook Cardiac Data). The results demonstrate that registration with DIRNet is as accurate as a conventional deformable image registration method with short execution times.

Keywords: Deep learning · Deformable image registration · Convolution neural network · Spatial transformer · Cardiac cine MRI

1 Introduction

Image registration is a fundamental step in many medical image analysis tasks. Traditionally, image registration is performed by exploiting intensity information between pairs of fixed and moving images. Since recently, deep learning approaches are used to aid image registration. Wu et al. [11] used a convolutional stacked auto-encoder (CAE) to extract features from fixed and moving images that are subsequently used in conventional deformable image registration algorithms. However, the CAE is decoupled from the image registration task and hence, it does not necessarily extract the features most descriptive for image registration. The training of the CAE was unsupervised, but the registration task was not learned end-to-end. Miao et al. [8] and Liao et al. [6] have used deep

M.J. Cardoso et al. (Eds.): DLMIA/ML-CDS 2017, LNCS 10553, pp. 204–212, 2017.
DOI: 10.1007/978-3-319-67558-9_24

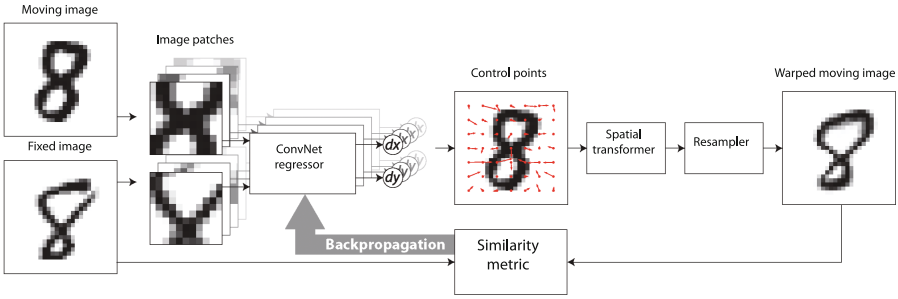

Fig. 1. Schematics of the DIRNet with two input images from the MNIST data. The DIRNet takes one or more pairs of moving and fixed images as its inputs. The fully convolutional ConvNet regressor analyzes spatially corresponding image patches from the moving and fixed images and generates a grid of control points for a B-spline transformer. The B-spline transformer generates a full displacement vector field to warp a moving image to a fixed image. Training of the DIRNet is unsupervised and end-to-end by backpropagating an image similarity metric as a loss.

learning to learn rigid registration with predefined registration examples. Miao et al. [8] used a convolutional neural network (ConvNet) to predict a transformation matrix for rigid registration of synthetic 2D to 3D images. Liao et al. [6] used a ConvNet for intra-patient rigid registration of CT to cone-beam CT applied to either cardiac or abdominal images. This ConvNet learned to predict iterative updates of registration using reinforcement learning. Both methods are end-to-end but use supervised techniques, i.e. registration examples are necessary for training, which are often task specific and highly challenging to obtain.

Jaderberg et al. [3] introduced the spatial transformer network (STN) that can be used as a building block that aligns input images in a larger network that performs a particular task. By training the entire network end-to-end, the embedded STN deduces optimal alignment for solving that specific task. However, alignment is not guaranteed, and it is only performed when required for the task of the entire network. The STNs were used for affine transformations, as well as deformable transformations using thin-plate splines. However, an STN needs many labeled training examples, and to the best of our knowledge, have not yet been used in medical imaging.

In this work, we present the deformable image registration network (DIRNet). The DIRNet takes pairs of fixed and moving images as inputs, and it outputs moving images warped to the fixed images. Training of the DIRNet is unsupervised. Unlike previous methods, the DIRNet is not trained with known registration examples, but learns to register images by directly optimizing a similarity metric between the fixed and the moving image. Hence, similar to conventional intensity-based image registration, it directly learns the registration task end-to-end and it is truly unsupervised. In addition, a trained DIRNet is able to perform deformable image registration non-iteratively on unseen data.

To the best of our knowledge, this is the first deep learning method for end-to-end unsupervised deformable image registration.

2 Method

The proposed DIRNet consists of a ConvNet regressor, a spatial transformer, and a resampler (Fig. 1). The ConvNet regressor analyzes spatially corresponding patches from a pair of fixed and moving input images and outputs local deformation parameters for the spatial transformer. The spatial transformer generates a dense displacement vector field (DVF) that enables the resampler to warp the moving image to the fixed image. The DIRNet learns the registration task end-to-end by unsupervised training with an image similarity metric. Since the training phase involves simultaneous optimization of registration of many image pairs, the ConvNet implicitly learns a representation of the features in images that are important for predictions of local displacement. Unlike regular image registration methods that typically perform iterative optimization for each image pair at hand, a trained DIRNet registers images in one pass.

The ConvNet regressor expects concatenated pairs of moving and fixed images as its input, and applies four alternating layers of 3×3 convolutions with 0-padding and 2×2 downsampling layers. Downsampling reduces the number of the ConvNet parameters, but it is associated with translational invariance. We postulate that this effect should be minimal in a ConvNet used for image registration, thus we use average pooling which should retain the most information during downsampling. Subsequently, 3×3 convolutional layers are added to increase the receptive field of the ConvNet to coincide with the capture range of the control points of the spatial transformer. Finally, three 1×1 convolutional layers are applied. Batch normalization [2] and exponential linear units [1] are used throughout the network, except in the final layer. The number of kernels per layer can be of arbitrary size, but the number of kernels of the final layer is determined by the dimensionality of the input images (e.g. 2 kernels for 2D images that require 2D displacement). The fully convolutional design in combination with downsampling allows fast analysis of separate spatially corresponding patch pairs from the fixed and moving images. The input image sizes and the number of downsampling layers jointly define the number of output parameters, i.e. the size and spacing of the control point grid. This way, for images of different sizes, similar grid spacing is ensured. Using the control point displacements, the spatial transformer generates a DVF used to warp the moving image to the fixed image. Like in [3], a thin-plate spline could be used as a spatial transformer, but due to its global support it is deemed less suitable for a patched-based approach. Therefore, we implemented a cubic B-spline [10] transformer which has local support. Thereafter, a resampler is used to generate warped moving images with linear interpolation.

The DIRNet is trained by optimizing an image similarity metric (i.e. by back-propagating dissimilarity) between pairs of moving and fixed images from a training set using mini-batch stochastic gradient descent (Adam [4]). Any similarity

metric used in conventional image registration could be used. In this work normalized cross correlation is employed.

After training, the DIRNet can be applied for registration of unseen image pairs from a separate dataset.

3 Data

The DIRNet was evaluated with handwritten digits from the MNIST database [5] and clinical MRI scans from the Sunnybrook Cardiac Data (SCD) [9].

The MNIST database contains 28×28 pixel grayscale images of handwritten digits that were centered by computing the center of mass of the pixels. The test images (10,000 digits) were kept separate from the training images (60,000 digits). One sixth of the training data was used for validation to monitor overfitting during training.

The SCD contains 45 cardiac cine MRI scans that were acquired on a single MRI-scanner. The scans consist of short-axis cardiac image slices each containing 20 timepoints that encompass the entire cardiac cycle. Slice thickness and spacing is 8 mm, and slice dimensions are 256×256 with a pixel size of $1.25\,\text{mm} \times 1.25\,\text{mm}$. The SCD is equally divided in 15 training scans (183 slices), 15 validation scans (168 slices), and 15 test scans (176 slices). An expert annotated the left ventricle myocardium at end-diastolic (ED) and end-systolic (ES) time points following the annotation protocol of the SCD. Annotations were made in the test scans and only used for final quantitative evaluation. In total, 129 image slices were annotated, i.e. 258 annotated timepoints.

4 Experiments and Results

DIRNet was implemented with Theano[1] and Lasagne[2], and conventional registration was performed with SimpleElastix [7].

4.1 Registration of Handwritten Digits

To demonstrate feasibility of the method we first applied it to registration of handwritten digits from the MNIST data. Separate DIRNet instances were trained for image registration of a specific class: one for each digit. The DIRNets were designed with 16 kernels per convolution layer, the third and fourth downsampling layers were removed. This resulted in a control point grid of 7×7 (grid spacing of 4 pixels). Each DIRNet was trained separately with random combinations of digits from its class with mini-batches of 32 random fixed and moving image pairs in 5,000 iterations (i.e. backpropagations). See Fig. 2 (left) for the learning curves.

[1] http://deeplearning.net/software/theano/ *(version 0.8.2).*
[2] https://lasagne.readthedocs.io/en/latest/ *(version 0.2.dev1).*

Registration performance of the trained DIRNets was qualitatively assessed on the test data. For each digit, one sample was randomly chosen to be the fixed image. Thereafter, all remaining digits (approximately 1,000 per class) were registered to the corresponding fixed image. Figure 2 (right) shows the registration results.

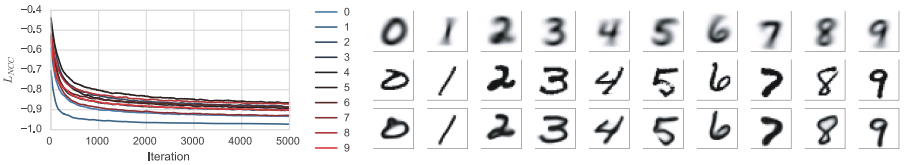

Fig. 2. Left: Learning curves showing the negative normalized cross correlation loss (L_{NCC}) on the validation set of DIRNets trained in 5,000 iterations for registration of MNIST digits. Right: Registration results of the trained DIRNets on a separate test set. The top row shows an average of all moving images per class (about 1,000 digits), the middle row shows one randomly chosen fixed image per class, and the bottom row shows an average of the registration results of independent registrations of the moving images to the chosen fixed image. Averages of the moving images after registration are much sharper than before registration indicating a good registration result. The blurry areas in the resulting images indicate where registration is challenging.

4.2 Registration of Cardiac MRI

Next, to demonstrate feasibility of the method on real-world medical data, we register cine cardiac MR images from the SCD. The DIRNet was trained by randomly selecting pairs of fixed and moving image slices from cardiac cine MRI scans (4D data). The pairs of fixed and moving images were anatomically corresponding slices from the same 4D scan of a single patient but acquired at different time points in the cardiac cycle. This resulted in 69,540 image pairs for training, and 63,840 pairs for validation.

A baseline DIRNet, as described in Sect. 2, was designed with 16 kernels per convolution layer. This resulted in a grid of 16×16 control points, i.e. a grid spacing of 16 pixels (20 mm). To evaluate effect of various DIRNet parameters, additional experiments were performed. First, to evaluate the effect of the downsampling method, DIRNet-A1 was designed with max-pooling layers, and DIRNet-A2 was designed with 2×2 strided convolutions. Second, to evaluate the effect of the spatial transformer, DIRNet-B1 was designed with a quadratic B-spline transformer, and DIRNet-B2 with a thin-plate spline transformer. Finally, to show the effect of the size of the receptive field (i.e. patch size), DIRNet-C1 was designed with neighbouring (i.e. non-overlapping) patches, by leaving out the last two 3×3 convolutional layers. In addition, DIRNet-C2 analyzed full image slices for each control point by replacing the 1×1 convolution layers with a 3×3 convolution layer, followed by a downsampling layer, two fully connected layers of 1,024 nodes, and a final output layer of 16×16 2D control points.

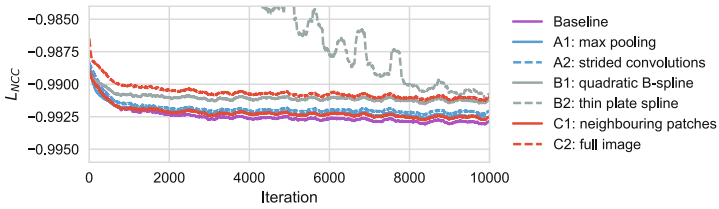

Fig. 3. Learning curves showing the negative normalized cross correlation loss (L_{NCC}) on the validation set of DIRNets The Net loss over 10,000 iteration for the baseline DIRNet, DIRNets with different downsampling techniques (A1, A2), DIRNets with different spatial transformers (B1, B2), and DIRNets with different receptive fields (C1, C2).

Table 1. Registration performance was quantified in cardiac MRI by registration of image slices and subsequent transformation of corresponding left ventricle annotations. Mean and standard deviations of the Dice coefficients between the reference and warped segmentations were computed. Additionally, 95$^{\text{th}}$ percentiles of the surface distance (95$^{\text{th}}$SD), and mean absolute surface distance (MAD) were calculated. The rows list results before registration, with conventional iterative image registration using SimpleElastix, and results obtained using the DIRNet. The rightmost column shows the runtime at inference for the conventional methods and the best performing DIRNet.

		Iterations	Dice	95$^{\text{th}}$ SD (mm)	MAD (mm)	Time (s)
No registration			0.62 ± 0.15	7.79 ± 2.92	2.89 ± 1.07	-
SimpleElastix		2×100	0.79 ± 0.08	5.09 ± 2.36	1.91 ± 0.94	0.51 ± 0.07
SimpleElastix		2×2000	$\mathbf{0.81 \pm 0.08}$	5.09 ± 7.25	$\mathbf{1.75 \pm 1.29}$	7.38 ± 0.94
DIRNet	BL		0.80 ± 0.08	$\mathbf{5.03 \pm 2.30}$	1.83 ± 0.89	0.049 ± 0.004
	A1		0.78 ± 0.08	5.26 ± 2.16	1.95 ± 0.85	-
	A2		0.78 ± 0.08	5.30 ± 2.28	1.97 ± 0.87	-
	B1		0.72 ± 0.11	6.41 ± 2.61	2.40 ± 0.96	-
	B2		0.78 ± 0.09	5.48 ± 2.36	2.01 ± 0.89	-
	C1		0.79 ± 0.08	5.20 ± 2.30	1.92 ± 0.89	-
	C2		0.76 ± 0.09	5.55 ± 2.24	2.10 ± 0.90	-

Each DIRNet was trained until convergence in mini-batches of 32 image pairs in at least 10,000 iterations. The training loss closely followed the validation loss in each experiment, and no signs of overfitting were apparent. Figure 3 shows the validation loss of 10,000 iterations during training for all experiments. The DIRNets converged quickly in each experiment, except DIRNet-B2, where convergence was reached after approximately 30,000 iterations. The final loss was lowest for baseline DIRNet.

Quantitative evaluation was performed on the test set by registering image slices at ED to ES, and vice versa, which resulted in 258 independent registration experiments. The obtained transformation parameters were used to warp

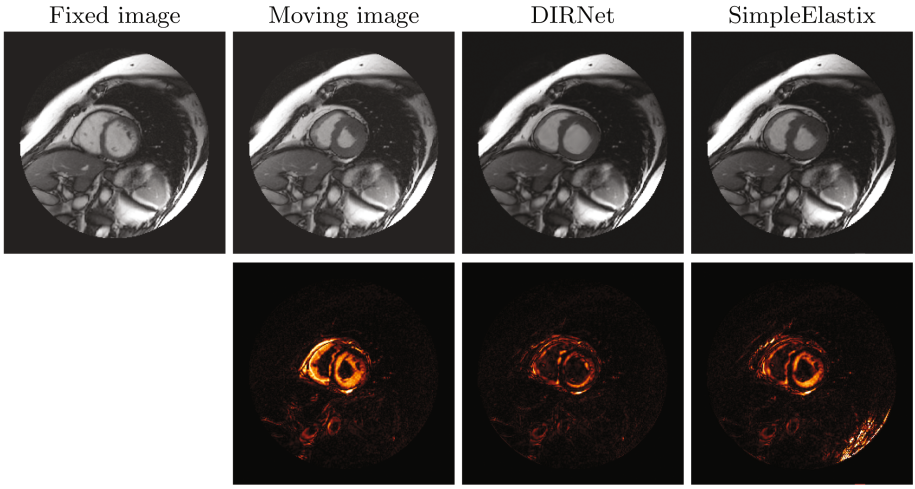

Fig. 4. Top, from left to right: The fixed (ED), the moving (ES), the DIRNet warped, and the SimpleElastix warped images. Bottom: Heatmaps showing absolute difference images between the fixed image and (from left to right) the original, the DIRNet warped, and the SimpleElastix warped moving images.

the left ventricle annotations of the moving image to the fixed image. The transformed annotations were compared with the reference annotations in the fixed images. The results are listed in Table 1. For comparison, the table also lists conventional iterative intensity-based registrations (SimpleElastix), with parameters specifically tuned for this task. A grid spacing was used similar to the DIRNet but in a multi-resolution approach, downsampling first with a factor of 2 and thereafter using the original resolution. Two conventional image registration experiments were performed, one for optimal speed (2 times 100 iterations), and one for optimal registration accuracy (2 times 2000 iterations), but at the cost of longer computation time. Experiments with the DIRNets were performed on an NVIDIA Titan X Maxwell GPU and experiments with SimpleElastix were performed on an Intel Xeon 1620-v3 3.5 GHz CPU using 8 threads. Figure 4 shows registration results for a randomly chosen image pair.

5 Discussion and Conclusion

A deep learning method for unsupervised end-to-end learning of deformable image registration has been presented. The method has been evaluated with registration of images with handwritten digits and image slices from cine cardiac MRI scans. The presented DIRNet achieves a performance that is as accurate as a conventional deformable image registration method with substantially shorter execution times. The method does not require training data, which is often difficult to obtain for medical images. To the best of our knowledge this is the first

deep learning method that uses unsupervised end-to-end training for deformable image registration.

Even though registration of images with handwritten digits is an easy task, the performed experiments demonstrate that a single DIRNet architecture can be used to perform registration in different image domains given domain specific training. It would be interesting to further investigate whether a single DIRNet instance can be trained for registration across different image domains.

Registration of slices from cardiac cine MRI scans was quantitatively evaluated between the ES and ED timepoints, so at maximum cardiac compression and maximum dilation. The conventional registration method (SimpleElastix) was specifically tuned for this task and the DIRNet was not, because it was trained for registration of slices from any timepoint of the cardiac cycle. Nevertheless, the results of the DIRNet were comparable to the conventional approach.

The data used in this work did not require pre-alignment of images. However, to extend the applicability of the proposed method in future work, performing affine registration will be investigated. Furthermore, proposed method is designed for registration of 2D images. In future work the method will be extended for registration of 3D images. Moreover, experiments were performed using only normalized cross correlation as a similarity metric, but any differentiable metric could be used.

To conclude, the DIRNet is able to learn image registration tasks in an unsupervised end-to-end fashion using an image similarity metric for optimization. Image registration is performed in one pass, thus non-iteratively. The results demonstrate that the network achieves a performance that is as accurate as a conventional deformable image registration method within shorter execution times.

Acknowledgments. This study was funded by the Netherlands Organization for Scientific Research (NWO): project 12726.

References

1. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (ELUs). In: ICLR (2016)
2. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: PMLR, vol. 37, pp. 448–456 (2015)
3. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 2017–2025. Curran Associates, Inc., Red Hook (2015)
4. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
5. LeCun, Y., Cortes, C.: The MNIST database of handwritten digits (1998)
6. Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., Comaniciu, D.: An artificial agent for robust image registration. arXiv preprint arXiv:1611.10336 (2016)

7. Marstal, K., Berendsen, F., Staring, M., Klein, S.: SimpleElastix: a user-friendly, multi-lingual library for medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2016)
8. Miao, S., Wang, Z.J., Liao, R.: A CNN regression approach for real-time 2D/3D registration. IEEE Trans. Med. Imaging **35**(5), 1352–1363 (2016)
9. Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A., Wright, G.: Evaluation framework for algorithms segmenting short axis cardiac MRI (2009)
10. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging **18**(8), 712–721 (1999)
11. Wu, G., Kim, M., Wang, Q., Munsell, B.C., Shen, D.: Scalable high performance image registration framework by unsupervised deep feature representations learning. IEEE Trans. Biomed. Eng. **63**(7), 1505–1516 (2016)