Deformable image registration using convolutional neural networks

Koen A. J. Eppenhof, Maxime W. Lafarge, Pim Moeskops, Mitko Veta, Josien P. W. Pluim

Koen A. J. Eppenhof, Maxime W. Lafarge, Pim Moeskops, Mitko Veta, Josien P. W. Pluim, "Deformable image registration using convolutional neural networks," Proc. SPIE 10574, Medical Imaging 2018: Image Processing, 105740S (2 March 2018); doi: 10.1117/12.2292443

SPIE.

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

Deformable image registration using convolutional neural networks

Koen A.J. Eppenhof, Maxime W. Lafarge, Pim Moeskops, Mitko Veta, and Josien P.W. Pluim

Medical Image Analysis Group (IMAGE/e), Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

ABSTRACT

Deformable image registration can be time-consuming and often needs extensive parameterization to perform well on a specific application. We present a step towards a registration framework based on a three-dimensional convolutional neural network. The network directly learns transformations between pairs of three-dimensional images. The outputs of the network are three maps for the x, y, and z components of a thin plate spline transformation grid. The network is trained on synthetic random transformations, which are applied to a small set of representative images for the desired application. Training therefore does not require manually annotated ground truth deformation information. The methodology is demonstrated on public data sets of inspiration-expiration lung CT image pairs, which come with annotated corresponding landmarks for evaluation of the registration accuracy. Advantages of this methodology are its fast registration times and its minimal parameterization.

Keywords: deformable image registration, machine learning, convolutional networks, thoracic CT

1. INTRODUCTION

A common problem with state-of-the-art deformable image registration algorithms is the amount of time required to optimize the cost function. Another problem is that every new application of a registration algorithm will require a specific parameter setting to achieve optimal performance. For example, registration algorithms that are designed to work well on high quality images of healthy patients are not guaranteed to work on images of lower quality or images containing pathology using the same parameter setting. These parameter settings are often adjusted manually, or not at all. To address these issues, we propose a novel registration approach based on fully convolutional neural networks. By turning image registration into a supervised problem, a registration algorithm can be trained in such a way that it is specifically optimized for a certain class of images, for example a certain type of pathology or a specific anatomy, taking away the need for manual parameterization. The proposed method estimates a transformation model for two input images directly from the images, resulting in a very fast registration algorithm.

1.1 Related work on machine learning in medical image registration

The application of machine learning techniques to image registrations has been studied in recent papers. The application areas include rigid 2D-to-3D registration, scoring the registration accuracy, and learning multimodal similarity metrics. The more recent papers include deep learning techniques like convolutional neural networks, which have been applied to a large number of other medical image analysis tasks such as segmentation, shape modeling, and detection tasks.[1] Previous work has employed machine learning techniques to directly perform rigid registration and deformable registration, to aid optimization of a similarity metric, to learn similarity metrics, and to validate image registration. Gouveia et al. compared multiple regression approaches for rigid 2D-to-3D registration approaches applied to simulated X-ray registration problems.[2] Miao et al. used convolutional neural networks to learn a regression of rigid registration parameters in 2D-to-3D registration.[3] Gutiérrez-Becker et al. developed a method that uses regression forests to learn multimodal motion predictors. These motion predictors were then used to estimate update steps for rigid deformable multimodal registration problems.[4]

Corresponding author: k.a.j.eppenhof@tue.nl

Medical Imaging 2018: Image Processing, edited by Elsa D. Angelini, Bennett A. Landman, Proc. of SPIE
Vol. 10574, 105740S · © 2018 SPIE · CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2292443

Muenzing et al. and Sokooti et al. developed methods that can learn to estimate registration quality metrics for non-linear registration based on classification and regression of the registration error respectively.[5,6] Eppenhof and Pluim developed a convolutional neural networks approach for regression of registration errors.[7] This method was trained on artificial transformations applied to a small set of training images. Simonovsky et al. and Wu et al. used convolutional neural networks to learn a similarity metric for multimodal image registration.[8,9]

Recent works have also attempted to use convolutional neural networks to estimate full displacement vector fields. Sokooti et al. used a convolutional neural network to estimate deformations from affinely registered pulmonary CT images.[10] De Vos et al. registered 2D MNIST data and 2D slices from 4D CTs using a convolutional neural network that was trained by backpropagating a similarity metric between the fixed and transformed moving images as measured by the normalized correlation coefficient.[11]

1.2 Aim of this paper

This paper aims to show that deformable transformations for 3D medical images can be estimated at very high speeds using convolutional neural networks. The network learns displacements on a thin plate spline grid that registers two 3D pulmonary images. Training does not require a manually annotated set of ground truth images, instead relying on learning synthetic transformations that are applied to a small set of representative images for the registration problem. Hence, no explicit choice for a similarity metric is required, because relevant features and metrics are implicitly learned from the data. In this paper we evaluate the network on a publicly available pulmonary data set, which is distinct from the training set, and comes with corresponding landmark annotations for computation of target registration errors.

2. METHODS

2.1 Transformation model

Let $I_F : \Omega_F \to \mathbb{R}$ and $I_M : \Omega_M \to \mathbb{R}$ be two real-valued images defined on their own d-dimensional spatial domains $\Omega_F \subset \mathbb{R}^d$ and $\Omega_M \subset \mathbb{R}^d$. Registration aims to find a transform $\mathbf{T} : \Omega_F \to \Omega_M$ between the domains of the fixed image I_F and the moving image I_M. We aim to find the transformation $\mathbf{T}$ for two three-dimensional images I_F and I_M using a convolutional neural network. The transform is defined as a thin plate spline (TPS)

$$\mathbf{T}(\mathbf{x}) = \mathbf{x} + A\mathbf{x} + \mathbf{t} + \sum_k \mathbf{c}_k \phi\left(\|\mathbf{d}_k\|\right) \tag{1}$$

where A is an affine matrix, $\mathbf{t}$ is a translation vector, and $\mathbf{c}_k$ are spline coefficients. The parameters $A, \mathbf{t}$, and $\mathbf{c}_k$ are computed from the displacements $\mathbf{d}_k$.

2.2 Network architecture

These displacements are defined on a grid covering the full image domain. The network's input are the two images I_F and I_M. The network's output consists of three maps, corresponding to the x, y, and z-components of the displacements $\mathbf{d}_k$. The network is based on a smaller version of the VGG-architecture[12] (Figure 1). Compared to the original VGG-implementation we have a pair of three-dimensional images as inputs instead of a single image, and instead of computing one output, the output consist of three maps corresponding to the x, y, and z components of a $6 \times 6 \times 6$ TPS grid. All convolutional layers use $3 \times 3 \times 3$ kernels with zero-padding to retain the size of the layer's input. The convolutions are followed by ReLU activation functions, except for the last $1 \times 1 \times 1$ convolutional layer, which has no activation function, allowing it to do regression of the grid components. Each convolutional layer is followed by a $2 \times 2 \times 2$ max-pooling layer that downsamples the input by a factor of two along each axis.

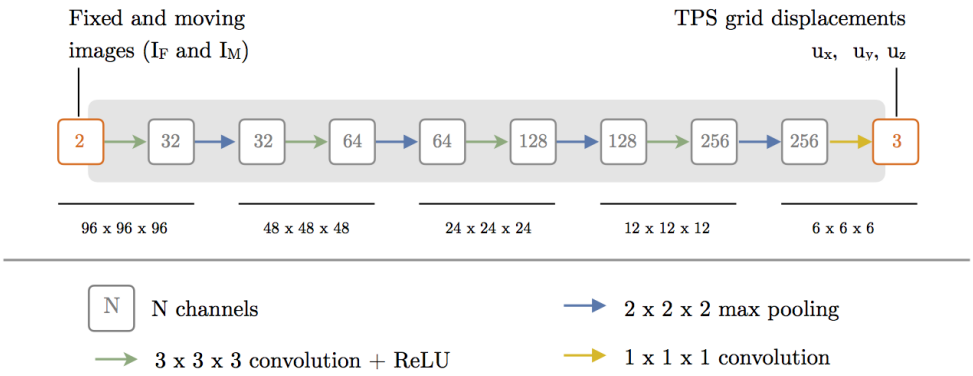

Figure 1. Network architecture. The network learns the displacements on a $6 \times 6 \times 6$ grid as three maps: one for every component of the thin plate spline grid. The input to the network is an image with two channels, where the channels correspond to the fixed and moving image of the registration.

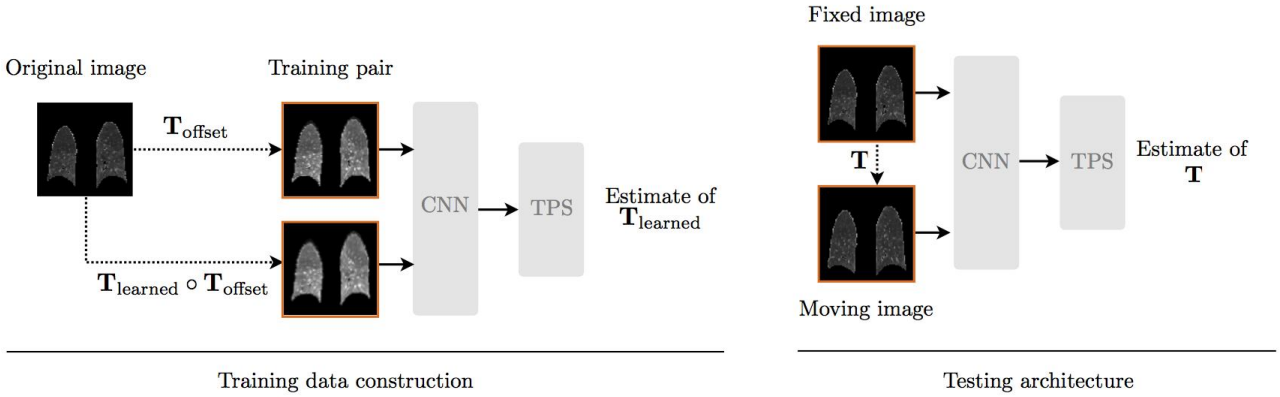

Figure 2. During training the inputs to the network are two transformed versions of the same image. During testing, the inputs are the fixed and moving images.

2.3 Training

The network was trained by minimizing the squared error of the estimated displacements $\hat{\mathbf{d}}_k$ averaged over the number of estimated vectors:

$$\text{MSE} = \frac{1}{N} \sum_{k=1}^{N} \|\mathbf{d}_k(\mathbf{x}) - \hat{\mathbf{d}}_k(\mathbf{x})\|^2. \tag{2}$$

This loss was optimized using stochastic gradient descent with a decaying learning rate

$$\eta(t) = \eta_0/(1 + \eta_0 \lambda t) \tag{3}$$

with $\eta_0 = 0.1$ and $\lambda = 10^{-4}$. We used single-instance batches (i.e. batch size = 1), and batch normalization on all convolutional layers with exponential moving averages of the normalization statistics over past iterations.[13] In the current implementation the network is trained on $96 \times 96 \times 96$ voxel images. For all experiments in this abstract the images are downsampled to this size using fourth-order B-spline interpolation before being used in the network. We found the current input size and architecture are a good trade-off between the amount of memory required and the accuracy of the network's estimates.

2.4 Training set

The training examples are constructed from a small set of images by applying synthetic transformations. For every iteration during training, a small random affine transformation is applied to an image $I(\mathbf{x})$ from the training

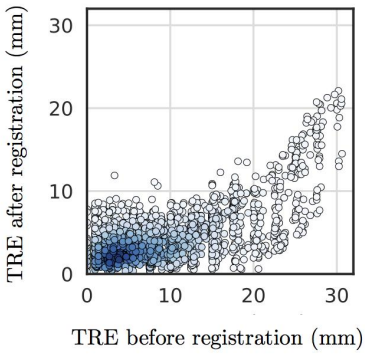

Figure 3. Post-registration TREs against pre-registration TREs. Every point corresponds to a landmark in the ten test images. Darker colors indicate a higher density of points.

set which results in an image $I(\mathbf{T}_{\text{offset}}(\mathbf{x}))$. We apply a larger transformation to the same image, composed of the offset transformation and a second, larger transformation $\mathbf{T}_{\text{learned}}$, which is actually learned by the network (Figure 2). The offset transformation serves as a form of data augmentation and is a purely affine transformation

$$\mathbf{T}_{\text{offset}}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} = (\mathbb{I} + \mathbf{B})\mathbf{x} + \mathbf{b}, \tag{4}$$

with $\mathbb{I}$ the 3D identity matrix, and the elements of $\mathbf{B}$ and $\mathbf{b}$ (in voxel units) sampled from a normal distribution $\mathcal{N}(0, 0.1)$. The learned transformation is modeled as a TPS transformation. The displacements $\mathbf{d}$ in Equation (1) are assigned to a $6 \times 6 \times 6$ grid that uniformly covers the image's domain. The magnitudes of the displacements are sampled from a three-dimensional uniform distribution. The range of this distribution needs to be set per application, based on the expected range of deformations. Combining the two transformations, every datum in the training set consists of a pair of images $I(\mathbf{T}_{\text{offset}}(\mathbf{x}))$ and $I((\mathbf{T}_{\text{learned}} \circ \mathbf{T}_{\text{offset}})(\mathbf{x}))$ with TPS transformation grid $\mathbf{T}_{\text{learned}}(\mathbf{x})$ as target. Both transforms are applied on-the-fly on an image from the small training set, resulting in unique inputs for every iteration of training.

2.5 Dataset

We used 3D thoracic CT images to train the network and validate our approach. These data come from two data sets: the DIRLAB set[14] being used for validation and the CREATIS data set for training.[15] The sets contain 10 and 7 pairs of 3D CT images respectively, showing the lungs at the end of inspiration and at the end of expiration. The DIRLAB CT data come with corresponding landmark annotations for the inspiration and expiration frames, with 300 landmarks per pair of images. Because the lungs move mainly upwards as a result of the breathing motion, the ranges of the displacements of the TPS transformation grid during training are set larger for the out-of-plane direction (d_z) compared to in-plane displacements (d_x and d_y). The displacements were sampled from uniform distributions with ranges of $d_x \in [-1, 1]$, $d_y \in [-1, 1]$, and $d_z \in [-1, 5]$ voxels. We compare our method's target registration error (TRE) to state-of-the art lung registration methods, which also were tested on the DIRLAB data set.

3. EXPERIMENTS

To test the network's ability to estimate deformations of the lungs, we ran it on the ten inspiration-expiration pairs of the DIRLAB data set. Figure 3 shows the improvement in TRE by our method. The graphs show a clear improvement for large deformations (larger than 10 mm prior to registration). Correlation plots for the x-, y-, and z-components of the vector field show a correlation between the ground truth and estimation, with Pearson correlation coefficients of 0.54, 0.65, and 0.80 for x, y, and z displacements respectively (Figure 4). To address the sliding motion of the ribs against the lung tissue, registration was limited to the lung fields extracted using lung masks. Target registration errors measured using the annotated landmarks and comparisons with other methods are shown in Table 1. Note that the compared algorithms all explicitly model the sliding motion

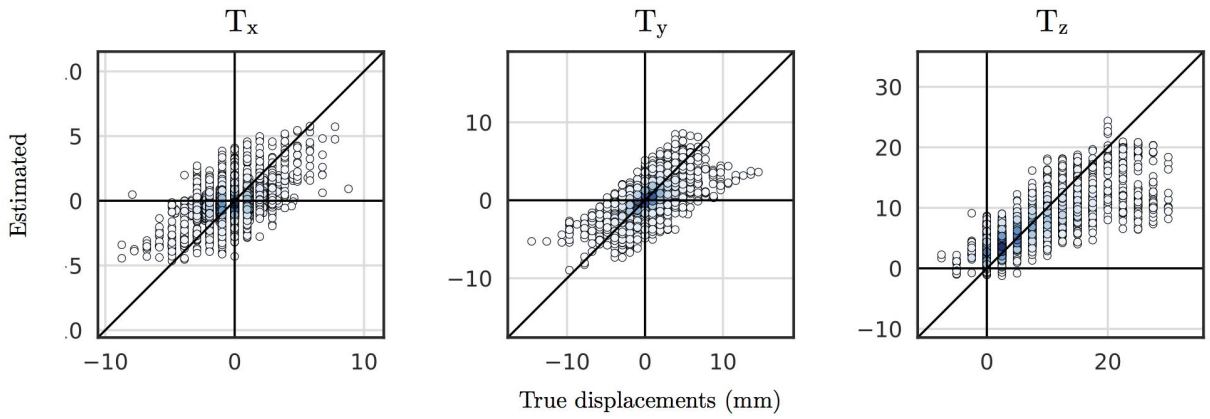

Figure 4. Correlation plots for each of the vector field components. Every point corresponds to a landmark in the ten test images. Darker colors indicate a higher density of points.

interface of the ribs and the lungs; something that is not included in our algorithm and which may contribute to larger TRE values. Estimation of the transformation took on average 55 ± 5 milliseconds on an NVIDIA GTX Titan X GPU compared to 100 minutes[16] for the method by Wu et al. and 58 minutes[16] for Delmon et al. Timing information for the methods by Schmidt-Richtberg et al. and Berendsen et al. is not reported in literature.

4. DISCUSSION

This paper describes an approach to estimate deformable image transformations directly from two images using convolutional neural networks for the purpose of 3D image registration. We have validated the method on CT inspiration-expiration pairs with corresponding landmark annotations. Our method shows that learning approaches are a viable approach for this registration problem. Especially the fast registration times make the method interesting for deformable registration problems where low latency is important, for example in radiation treatment. Other advantages are that manually annotation is not necessary to train the network. We have shown that a simple form of data augmentation allows training on a very small set of images, and can generalize to a different data set of CT images acquired from different scanners and patient groups. By using artificial deformations, only a set of images that is similar to the target application is required to train the network. Making these deformations more realistic is an important topic for future research. Two limitations of the current implementation are the limited input size, which is likely the largest cause of registration errors, and the inability of the network to estimate large displacements, as the TRE reduction plot in Figure 3 shows. We hypothesize that accurate estimation of a larger range of errors requires multi-stage or multi-resolution strategies. A possible solution is training multiple networks to perform a sequence of registrations at different resolutions or with different transformation models. In our experiments we found that estimating the correct range of the displacements for the training transformation is very important when training. Future work will aim to further improve the accuracy of the method, by making the deformations in the training set more realistic for pulmonary images, by training on larger images, and by using a more fine-grained transformation model.

REFERENCES

[1] Greenspan, H., van Ginneken, B., and Summers, R. M., "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE TMI* **35**(5), 1153–1159 (2016).

[2] Gouveia, A. I. R., Metz, C., Freire, L., and Klein, S., "Comparative evaluation of regression methods for 3D-2D image registration," in [*Proc. ICANN*], 238–245 (2012).

[3] Miao, S., Wang, Z. J., and Liao, R., "A CNN regression approach for real-time 2D/3D registration," *IEEE TMI* **35**(5), 1352–1363 (2016).

[4] Gutiérrez-Becker, B., Mateus, D., Peter, L., and Navab, N., "Learning optimization updates for multimodal registration," in [*Proc. MICCAI*], 19–27 (2016).

Table 1. Target registration errors in millimeters measured on annotated corresponding landmarks, comparing our method with state-of-the-art methods for lung registration. Standard deviations over all 300 landmarks in parentheses.

Image	No registration	Schmidt-Richtberg et al.[17]	Wu et al.[18]	Delmon et al.[16]	Berendsen et al.[19]	Proposed
1	3.89 (2.78)	1.22 (0.64)	1.1 (0.5)	1.2 (0.6)	1.00 (0.52)	1.65 (0.89)
2	4.34 (3.90)	1.14 (0.65)	1.0 (0.5)	1.1 (0.6)	1.02 (0.57)	2.26 (1.16)
3	6.94 (4.05)	1.36 (0.81)	1.3 (0.7)	1.6 (0.9)	1.14 (0.89)	3.15 (1.63)
4	9.83 (4.86)	2.68 (2.79)	1.5 (1.0)	1.6 (1.1)	1.46 (0.96)	4.24 (2.69)
5	7.48 (5.51)	1.57 (1.23)	1.9 (1.5)	2.0 (1.6)	1.61 (1.48)	3.52 (2.23)
6	10.9 (6.97)	2.21 (1.66)	1.6 (0.9)	1.7 (1.0)	1.42 (0.89)	3.19 (1.50)
7	11.0 (7.43)	3.81 (3.06)	1.7 (1.1)	1.9 (1.2)	1.49 (1.06)	4.25 (2.08)
8	15.0 (9.01)	3.42 (4.25)	1.6 (1.4)	2.2 (2.3)	1.62 (1.71)	9.03 (5.08)
9	7.92 (3.98)	1.83 (1.19)	1.4 (0.8)	1.6 (0.9)	1.30 (0.76)	3.85 (1.86)
10	7.30 (6.35)	2.06 (1.92)	1.6 (1.2)	1.7 (1.2)	1.50 (1.31)	5.07 (2.31)
mean	8.46 (5.48)	2.13 (1.82)	1.47 (0.96)	1.66 (1.14)	1.36 (1.01)	4.02(3.08)

[5] Muenzing, S. E. A., van Ginneken, B., Murphy, K., and Pluim, J. P. W., "Supervised quality assessment of medical image registration: Application to intra-patient CT lung registration," *Medical Image Analysis* **16**(8), 1521–1531 (2012).

[6] Sokooti, H., Saygili, G., Glocker, B., Lelieveldt, B. P. F., and Staring, M., "Accuracy estimation for medical image registration using regression forests," in [*Proc. MICCAI*], 107–115 (2016).

[7] Eppenhof, K. A. J. and Pluim, J. P. W., "Supervised local error estimation for nonlinear image registration using convolutional neural networks," in [*Proc. SPIE Medical Imaging*], **10133**, 101331U (2017).

[8] Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., and Komodakis, N., "A deep metric for multimodal registration," in [*Proc. MICCAI*], 10–18 (2016).

[9] Wu, G., Kim, M., Wang, Q., Munsell, B. C., and Shen, D., "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE TMI* **63**(7), 1505–1516 (2016).

[10] Sokooti, H., Vos, B. D. D., Berendsen, F. F., Lelieveldt, B. P. F., Isgum, I., and Staring, M., "Nonrigid image registration using multi-scale 3D convolutional neural networks," in [*Proc. MICCAI*], 232–239 (2017).

[11] Vos, B. D. D., Berendsen, F. F., Viergever, M. A., Staring, M., and Isgum, I., "End-to-end unsupervised deformable image registration with a convolutional neural network," in [*Proc. DLMIA workshop held in conjunction with MICCAI*], 204–212 (2017).

[12] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," in [*Proc. ICLR*], (2015).

[13] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in [*Proc. ICML*], 448–456 (2015).

[14] Castillo, E., Castillo, R., Martinez, J., Shenoy, M., and Guerrero, T., "Four-dimensional deformable image registration using trajectory modeling," *Physics in Medicine and Biology* **55**(1), 305–327 (2009).

[15] Vandemeulebroucke, J., Bernard, O., Rit, S., Kybic, J., Clarysse, P., and Sarrut, D., "Automated segmentation of a motion mask to preserve sliding motion in deformable registration of thoracic CT," *Medical Physics* **39**(2), 1006–1015 (2012).

[16] Delmon, V., Rit, S., Pinho, R., and Sarrut, D., "Registration of sliding objects using direction dependent B-splines decomposition," *Physics in Medicine and Biology* **58**(5), 1303–1314 (2013).

[17] Schmidt-Richberg, A., Werner, R., Handels, H., J., and Ehrhardt, "Estimation of slipping organ motion by registration with direction-dependent regularization," *Medical Image Analysis* **16**(1), 150–159 (2012).

[18] Wu, Z., Rietzel, E., Boldea, V., Sarrut, D., and Sharp, G. C., "Evaluation of deformable registration of patient lung 4DCT with subanatomical region segmentations," *Medical Physics* **35**(2), 775–781 (2008).

[19] Berendsen, F. F., Kotte, A. N. T. J., Viergever, M. A., and Pluim, J. P., "Registration of organs with sliding interfaces and changing topologies," in [*Proc. SPIE Medical Imaging*], 90340E–1 (2014).