






Learning a Deformable Registration Pyramid

Niklas Gunnarsson^{1,2} , Jens Sjölund^{1,2} , and Thomas B. Schön¹ 

¹ Department of Information Technology, Uppsala University, Uppsala, Sweden
{niklas.gunnarsson,jens.sjolund,thomas.schon}@it.uu.se

² Elekta Instrument AB, Stockholm, Sweden
{niklas.gunnarsson,jens.sjolund}@elekta.com

Abstract. We introduce an end-to-end unsupervised (or weakly supervised) image registration method that blends conventional medical image registration with contemporary deep learning techniques from computer vision. Our method downsamples both the fixed and the moving images into multiple feature map levels where a displacement field is estimated at each level and then further refined throughout the network. We train and test our model on three different datasets. In comparison with the initial registrations we find an improved performance using our model, yet we expect it would improve further if the model was fine-tuned for each task. The implementation is publicly available (<https://github.com/ngunnar/learning-a-deformable-registration-pyramid>).

Keywords: Medical image registration · Deep learning · Deformable registration.

1 Introduction

Image registration is a fundamental problem in medical imaging. It is widely used in applications to, for example, combine images of the same object from different modalities (multimodal registration), detect changes between images at different times (spatiotemporal registration), and map segments from a predefined image to a new image (atlas based segmentation).

The basic principle of image registration is to find a displacement field ϕ that maps positions in a moving image to the corresponding positions in a fixed image. Conventionally, image registration problems are often stated as optimization problems, where the aim is to minimize a complex energy function [1].

A popular heuristic for solving image registration problems is to use a coarse-to-fine approach [2] i.e. to start with a rough estimate of the displacement field and refine it in one or several steps. It is common to downsample the fixed and moving images using a kernel based pyramid, and make a first estimate of the displacement field at the lowest resolution which is then used as an initial guess when estimating the field at the next resolution level, and so forth.

Due to the complexity of the energy function each estimate is computationally expensive and requires long execution time. Machine learning provides an alternative approach, where a model is optimized (learned) offline based on a training dataset, obviating the need for expensive optimization at test time [3]. In this paper we present an image registration method that combines the conventional coarse-to-fine approach with a convolutional neural network (CNN).

2 Method

We have developed a 3D deformable image registration method inspired by the PWC-Net [4], a 2D optical flow method popular in computer vision. Our method estimates and refines a displacement field at each level of a CNN downsampling pyramid.

2.1 Architecture

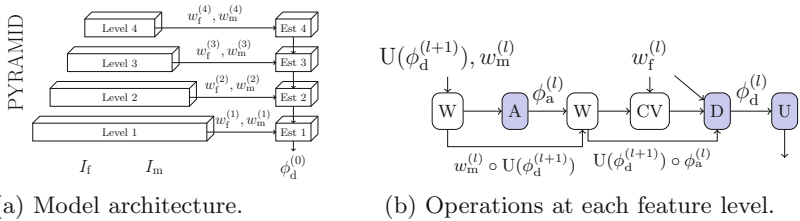


Fig. 1. An overview of the model architecture. The moving and fixed image are downsampled into several feature maps using the pyramid (a). Figure (b) shows operations at each feature level. Blue and white boxes represent operations with and without trainable parameters, respectively.

The pyramid downsamples the moving image I_m and the fixed image I_f into several feature maps $\{w_m^{(l)}, w_f^{(l)}\}_{l=1}^L$. At each level, starting from the top, a displacement field $\phi_d^{(l)}$ is estimated and used as an initial guess at finer levels. Figure 1 illustrates the model architecture (a) and operations at each level (b). The total number of trainable parameters in our model is 8.6 million. Our model includes multiple CNN blocks. These consist of a 3D convolutional layer followed by Leaky Relu and batch normalization. All 3D convolutional layers use a kernel size of (3,3,3). Each module of our model is explained below:

Pyramid: Downsamples the moving and fixed image into several feature map levels using 3D CNN layers. The same pyramid is used for the moving and the fixed images. We use a four-level pyramid ($L = 4$) where each level consists of three CNN blocks. The stride is two in the first block and one in the subsequent blocks. The number of filters at each level is 16, 32, 32, and 32, respectively.

Warp (W): Warps features from moving images with the estimated displacement field. This module has no trainable parameters.

Affine (A): A dense neural network that estimates the 12 parameters in an affine transformation. This module consists of a global average pooling followed by a dense layer.

Cost volume (CV): Correlation between the warped feature maps from the moving image and feature maps from the fixed image. For computational reasons the cost volume is restricted to voxel neighborhoods of size d . This module has no trainable parameters.

Deform (D): A 3D DenseNet [5] that estimates the displacement field based on its current estimate, the cost volume and the feature maps from the fixed image. This module uses 5 CNN blocks of the same type as in the Pyramid but with 64, 64, 32, 18, and 8 filters, respectively followed by a convolutional layer with 3 filters.

Upsample (U): Upsamples the estimated displacement field from one level to the next. Consists of an upsampling layer followed by a single 3D CNN.

2.2 Loss Function

Our loss function combines image similarity with regularization of the displacement field. By including the intermediate estimates in the loss, we aim to gain additional control of the network. Auxiliary information, e.g. anatomical segmentations S_m and S_f are incorporated via an additional structural similarity term \mathcal{L}_{seg} . Our resulting loss function can be written as

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \sum_{l=0}^L \left(\mathcal{L}_{\text{sim}}^{(l)} + \mathcal{L}_{\text{smooth}}^{(l)} \right). \quad (1)$$

We use the (soft) Dice coefficient (DCS) [6] for structural similarity and the normalized cross-correlation (NCC) [7] for image similarity. To ensure smooth displacements we regularize the affine displacement field with the L2-loss between the estimated value and an identity displacement field ($\phi_0^{(l)}$) and the deformable field with the spatial gradient of the displacement field [8],

$$\mathcal{L}_{\text{seg}} \left(S_f, S_m, \phi_d^{(0)} \right) = \lambda (1 - \text{DCS}(S_f, S_m \circ \phi_d^{(0)})), \quad (2a)$$

$$\mathcal{L}_{\text{sim}}^{(l)} \left(I_f^{(l)}, I_m^{(l)}, \phi_d^{(l)} \right) = -\gamma^{(l)} \text{NCC}(I_f^{(l)}, I_m^{(l)} \circ \phi_d^{(l)}), \quad (2b)$$

$$\mathcal{L}_{\text{smooth}}^{(l)} \left(\phi_a^{(l)}, \phi_d^{(l)} \right) = \alpha^{(l)} \|\phi_a^{(l)} - \phi_0^{(l)}\|_2^2 + \beta^{(l)} \|\nabla \phi_d^{(l)}\|_2^2, \quad (2c)$$

where $I_m^{(l)}$ and $I_f^{(l)}$ represent downsampled versions of the moving and fixed images at each level and $\phi_a^{(l)}$ and $\phi_d^{(l)}$ indicate the estimated affine and deformable registrations (for each level). The hyperparameters $\lambda, \{\gamma^{(l)}, \alpha^{(l)} \text{ and } \beta^{(l)}\}_{l=0}^L$ determine the importance of the corresponding terms.

3 Experiment

We evaluated the model on three different tasks from the 2020 Learn2Reg challenge [9]. The different tasks were: inspiration and expiration CT scans of thorax images with automatic segmented lung (Task 2) [10]; 3D CT abdominal images with thirteen segmented organs (Task 3); and segmented hippocampus MRI of healthy adults and adults with non-affective psychotic disorder (Task 4) [11].

We trained our model on image pairs from all tasks at the same time. All images were downsampled (to a resolution of $64 \times 64 \times 64$) and normalized ($I_f, I_m \in [0, 1]$). The different hyperparameters were $\lambda = 5.0$, $\gamma^{(l)} = 5/2^l$, $\alpha^{(l)} = 2^l$ and $\beta^{(l)} = 1/2^l$ for $l \in \{0, \dots, 4\}$ and for cost volume search range we used $d = 2$. The network was trained end-to-end using the Adam optimizer and a learning rate of 10^{-4} . To speed up training we used distributed training on three Nvidia GeForce GTX 1080 Ti graphic cards and trained the model for 100 epochs, which took approximately 24 h.

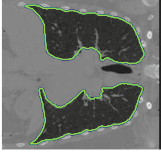
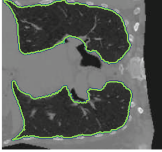
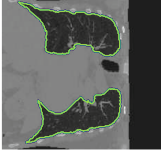
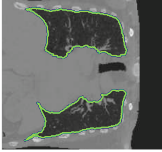
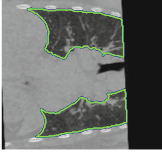
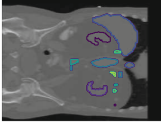
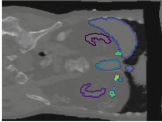
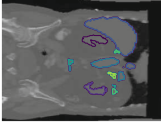
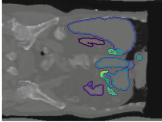
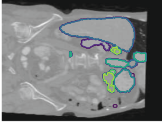
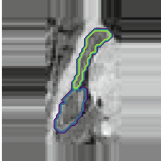
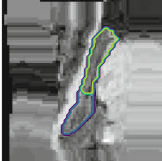

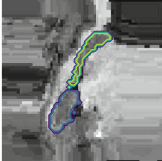
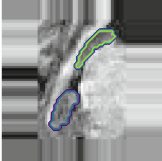
The results are shown in Table 1. Table 2 shows examples of warping the moving image using displacement fields $\phi_d^{(l)}$ estimated at three different levels $l \in \{0, 2, 4\}$. Based on the total score, our approach was ranked 5th according to the public leaderboard [9].

Table 1. Result on test dataset for each task.

Task	Method	TRE [12]	TRE30	DCS [13]	DCS30	HD95 [14]	SDlogJ [15]	Time (s)	
								GPU	CPU ^a
2	Our	9.00	12.22	–	–	–	0.12	0.31	4.83
	Initial	10.24	17.77	–	–	–	0.00	–	–
3	Our	–	–	0.39	0.12	43.03	0.13	0.31	4.83
	Initial	–	–	0.23	0.01	46.07	0.00	–	–
4	Our	–	–	0.74	0.67	2.82	0.16	0.32	4.83
	Initial	–	–	0.55	0.36	3.91	0.00	–	–

^a Prediction time only, excluding pre - and post processing.

Table 2. Sample result from the validation dataset. The moving image I_m (left) is warped with the estimated displacement field from several levels ($l = 4, 2, 0$). Starting from the coarsest to the finest level. The fixed image I_f is shown to the right.

	I_m	$I_m \circ \phi_d^{(4)}$	$I_m \circ \phi_d^{(2)}$	$I_m \circ \phi_d^{(0)}$	I_f
Task 2					
Task 3					
Task 4					

4 Conclusion and Future Work

In this paper we have shown that it is possible to include domain knowledge when developing machine learning methods for medical image registration problems. Our method operates in a coarse-to-fine manner and could be modified in many ways, e.g. by replacing the CNN pyramid with other technologies; like a Laplacian pyramid, similar to the winner of the competition [16], or modifying/removing displacement fields estimations (affine or deformable) in the levels.

In comparison with other participants in the competition our approach was to create a single general model for all tasks while other participants used different models or different training procedures [16–18] for each task. The general approach showed increased performance compared with initial registrations. In future work, we will evaluate to what extent the performance improves when fine tuning the model for each task.

During the training phase the memory usage was high (11.4 GB). In the experiments we downsampled the input images to a low resolution, using a batch size of one (at each GPU replica) and our partial cost volume had a search range of two to be able to fit the model in GPU memory (11.7 GB). We believe that an in-depth analysis of the network will reveal ways of reducing memory usage without sacrificing performance substantially, e.g. by removing superfluous layers or reducing the number of filters. One idea is to reduce the number of parameters in the DenseNet [19]. Other potential improvements include: 1) training each level separately, starting from the coarsest, which will reduce the number of

trainable parameters in each training process, 2) training the model on slices (2D) or thin slabs (2.5D), instead of the entire volume and iteratively estimate the entire 3D displacement field.

Acknowledgement. This research was funded by the *Wallenberg AI, Autonomous Systems and Software Program (WASP)* funded by Knut and Alice Wallenberg Foundation, and the Swedish Foundation for Strategic Research grant SM19-0029.

References

1. Hill, D.L.G., Batchelor, P.G., Holden, M., Hawkes, D.J.: Medical image registration. *Phys. Med. Biol.* **46**(3), R1 (2001)
2. Zitova, B., Flusser, J.: Image registration methods: a survey. *Image Vis. Comput.* **21**(11), 977–1000 (2003)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging* **38**(8), 1788–1800 (2019)
4. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943 (2018)
5. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708 (2017)
6. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571. IEEE (2016)
7. Pratt, W.K.: Digital image processing, 4th edition. *J. Electron. Imaging* **16**(2), 29901 (2007)
8. Estienne, T., et al.: U-ReSNet: ultimate coupling of registration and segmentation with deep nets. In: Shen, D., et al. (eds.) *MICCAI 2019*. LNCS, vol. 11766, pp. 310–319. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_35
9. Dalca, A., et al.: Learn2reg - the challenge (2020). <https://learn2reg.grand-challenge.org/>, <https://doi.org/10.5281/ZENODO.3715652>
10. Hering, A., Murphy, K., van Ginneken, B.: Lean2Reg challenge: CT lung registration - training data, May 2020
11. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019)
12. Fitzpatrick, J.M., West, J.B., Maurer, C.R.: Predicting error in rigid-body point-based registration. *IEEE Trans. Med. Imaging* **17**(5), 694–702 (1998)
13. Dice, L.R.: Measures of the amount of ecologic association between species. *Ecology* **26**(3), 297–302 (1945)
14. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(9), 850–863 (1993)
15. Kabus, S., Klinder, T., Murphy, K., van Ginneken, B., Lorenz, C., Pluim, J.P.W.: Evaluation of 4D-CT lung registration. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) *MICCAI 2009*. LNCS, vol. 5761, pp. 747–754. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04268-3_92

16. Mok, T.C.W., Chung, A.C.S.: Large deformation diffeomorphic image registration with Laplacian pyramid networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 211–221. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_21
17. Heinrich, M.P.: Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 50–58. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_6
18. Estienne, T.: Deep learning based registration using spatial gradients and noisy segmentation labels. arXiv preprint [arXiv:2010.10897](https://arxiv.org/abs/2010.10897) (2020)
19. Jégou, S., Drozdal, M., Vazquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: fully convolutional DenseNets for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 11–19 (2017)