Robust Unsupervised Image to Template Registration Without Image Similarity Loss

Slim Hachicha, Célia Le, Valentine Wargnier-Dauchelle, Michaël Sdika

Robust Unsupervised Image to Template Registration Without Image Similarity Loss

Slim Hachicha*[1], Célia Le*[1], Valentine Wargnier-Dauchelle[1], and Michaël Sdika[1]

Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, LYON, France

This preprint has not undergone peer review or any post-submission improvements or corrections. The Version of Record of this contribution will be published as a proceeding of The 2nd Workshop of Medical Image Learning with Limited & Noisy Data (MILLanD) held in conjunction with the 26th International Conference on Medical Image Computing & Computer Assisted Intervention (MICCAI 2023), Canada

Abstract. Although a giant step forward has been made in medical images analysis thanks to deep learning, good results still require a lot of tedious and costly annotations. For image registration, unsupervised methods usually consider the training of a network using classical registration dissimilarity metrics. In this paper, we focus on the case of affine registration and show that this approach is not robust when the transform to estimate is large. We propose an unsupervised method for the training of an affine image registration network without using dissimilarity metrics and show that we are able to robustly register images even when the field of view is significantly different in the image.

Keywords: image registration · unsupervised

1 Introduction

Image registration consists in finding a geometrical transformation to reposition an image, the moving image, in the spatial coordinate system of another image, the fixed image [5]. It has many applications such as longitudinal studies, studies on lung or cardiac motion or spatial normalization. In classical registration, an optimization problem is solved to minimize the dissimilarity between the fixed image and the moving image warped with the estimated transform [13]. The dissimilarity metric can be either geometrical or based on raw image intensities such as mutual information (MI) [9] to avoid the potentially unreliable and/or tedious geometric features extraction step. Typical problem of these optimization

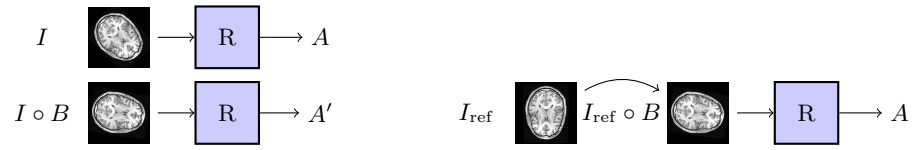
* The first two authors contributed equally to this work

based methods is the lack of robustness to artifacts, to bad initialization or to the presence of abnormality as well as their long computation time.

Deep learning registration methods have also been investigated to cope with these limitations. Supervised methods [8, 11, 10] are indeed robust and efficient but require the ground truth transformation to be known for images in the training set. As ground truth transforms are often impossible to obtain, the training is done with either (potentially unrealistic) synthetic transforms, estimated transforms from (potentially unreliable) third party registration and sometimes using additional annotations (that can be tedious to get). In [7, 1, 14], the deep registration framework is unsupervised: classical registration dissimilarity losses such as MI are used to train the network. Most limits of classical registration are re-introduced. These issues are also present when the dissimilarity loss is learnt as in [2]. Most deep registration methods also use, as input tensor, the moving and fixed image concatenated in different channels. This approach is problematic as, for large displacements, first fine scale convolution layers of the network will attempt to create features from unrelated parts of the pair of images. This defect can be alleviated by using each image as an independent input of two networks as in [4]: a network estimates the position of some keypoints for each image. An affine transform is then fitted to make the positions of the keypoints match. The network is first trained using supervised training with synthetic affine transforms but then trained on real pairs of images with dissimilarity losses.

As far as we know, only in [12] is addressed the problem of unsupervised registration without image dissimilarity losses: the fixed and moving image are encoded with a separate encoder, a correlation matrix between local features is computed providing a rough displacement likelihood map for each cell and a robust fit finally outputs the estimated affine transformation matrix. Only a cycle consistency loss is used for the training. However, the computation of this correlation matrix and the subsequent fit can be computationally demanding, it cannot scale to deformable registration for example. Furthermore, as the correlation is computed with the features of the moving image before warping, these features need to be affine invariant: affine equivariant features would be more discriminative for registration.

In this paper, we propose a deep unsupervised registration method. The network only takes the moving image as input and directly outputs the affine transform to register it to a reference template. Pairwise registration can then be easily obtained by composition. For training, only unregistered images without any label are required and dissimilarity losses are not used. Our method is robust to the presence of strong artefacts or abnormalities and is robust to extreme rotations, translations, scaling or shearing. The main contributions are: 1/ an unsupervised image registration method to train a network that directly outputs the affine transform of an image to an atlas 2/ a two steps training procedure to enlarge the range of affine matrices that can be estimated to the most extreme cases, 3/ a numerical evaluation showing the robustness of our approach even in the presence of large crops or occlusions.



(a) Equivariance loss: $\|BA' - A\|^2$ (b) Reference Loss: $\|BA - I_d\|^2$

Fig. 1: Unsupervised registration equivariance and reference loss

2 Method

2.1 Unsupervised Registration to a Reference Template without Dissimilarity Losses

Our goal is to train a neural network R whose input is an image I and the output is the transformation A such that $I \circ A$ is in the reference coordinate system. In other words, $I \circ R(I)$ is in the reference coordinate system. It is assumed that only a single image I_{ref} is in the reference coordinate system : $R(I_{\text{ref}}) = I_d$, all other images are neither registered nor labeled in any way. Pairwise registration is a straightforward consequence of template registration: once R is trained, the transform between a fixed image I_{fix} and a moving image I_{mov} is $R(I_{\text{mov}}) \circ [R(I_{\text{fix}})]^{-1}$.

Equivariance as an unsupervised registration loss As the network R registers all input images to the reference coordinate system, it should be affine equivariant. Indeed, for any input image I and transform B, the network should reposition $I \circ B$ in the reference coordinate system: $(I \circ B) \circ R(I \circ B)$ should always be registered. As $(I \circ B) \circ R(I \circ B) = I \circ (B \circ R(I \circ B)) = I \circ R(I)$, this means that for any image I and transform B, we should have:

$$B \circ R(I \circ B) = R(I). \tag{1}$$

This is a very strong structural prior of the task to solve, that is traditionally not enforced when learning registration networks. To enforce this property, for each image I of the training dataset, we draw a random affine transform B, and use the following equivariance loss:

$$L_{\text{equiv}} = \|B \circ R(I \circ B) - R(I)\|^2 \tag{2}$$

where the norm can by any matrix norm. Note that this is similar to the cycle consistency loss used in [12] for pairwise registration where their affine matrix is fitted from a correlation matrix.

The loss in eq. 2 itself is not sufficient to train the network. First, at no point so far, the reference coordinate system was specified: at an extreme limit, equivariance could be enforced for each image independently. Second, an identically null network R will satisfy eq. 1: some sort of regularization is required.

Regularization loss The regularization aims at discarding unrealistic affine transforms from the output of the neural network while retaining potentially large possible transforms. This loss L_{reg} is composed of two terms:

$$L_{\text{size}} = relu\left(\frac{1}{K} - \sigma\right) + relu(\sigma - K) \tag{3}$$

where σ are the singular values of the affine matrix predicted by the network and K is a scale hyperparameter and

$$L_{\text{anis}} = \|\log(\sigma_{\min}) - \log(\sigma_{\max})\|^2 \tag{4}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are respectively the minimal and maximal singular value. The first term penalizes extreme size variations, the second term aims to avoid a too strong anisotropy. Translation or rotation are not penalized.

Reference loss We consider the unique reference image of the training dataset I_{ref}: $R(I_{\text{ref}}) = I_d$. For any affine transform B, eq. 1 becomes $B \circ R(I_{\text{ref}} \circ B) = I_d$. Our reference loss is then (with any matrix norm):

$$L_{\text{ref}} = \|B \circ R(I \circ B) - Id\|^2, \tag{5}$$

where B is an affine transform randomly drawn during training.

One point is noticeable in our framework: at no point we need to compare pairs of images during the training. The reference image is fed through the network independently of all other images of the training dataset. No comparison is done between the warped image and the reference with a dissimilarity loss. No fit, that may potentially fail in some cases, need to be done between some sort of features of the moving and fixed images.

2.2 Increasing the Registration Range

Difference of field of view between the fixed and the moving images is a common registration failure reason. In order to increase the range of transformations our network is able to register, we propose a two-step training procedure. The network is first trained using our unsupervised framework on an unregistered training images that required moderate transformations to be registered. Once convergence is reached, this network is used to replace all the training images in the reference space. As all images are now registered, the network can then be trained using the loss L_{ref} with large affine parameters range on all the images of the training dataset. In the following, this training procedure will be referred as "two steps" method in opposition to the "one step" method without the supervised training with self registered images.

	D_0	D_{trans}	D_{shear}	D_{scale}	D_1
Translation	±10 pixels	±54 pixels	±10 pixels	±10 pixels	±54 pixels
Rotation	$\pm\ \pi$	$\pm\ \pi$	$\pm\ \pi$	$\pm\ \pi$	$\pm\ \pi$
Shearing	±10%	±10%	±50%	±10%	±50%
Scaling	±10%	±10%	±10%	±50%	±50%

Table 1: Affine transformation parameters range

3 Experiments

3.1 Material & Methods

Image Data The development of the algorithms was made using T2 brain images of the HCP database (www.humanconnectome.org/study/hcp-young-adult) that we split into 500 training subjects, 100 validation subjects, and 500 testing subjects. HCP images are provided registered together, we resample them to the 2mm T1 MNI atlas that we used as the I_{ref} image. To simplify and shorten the development, the middle axial slice was extracted, resulting in a dataset of 2D 109x91 registered brain MRIs but nothing in the method limits its application to 3D images.

Evaluation protocol To mimic an unregistered dataset, a random affine transformation is applied to each image before it goes into the network. The said transformation is considered unknown during the training phase but known for the evaluation reporting. It also allows to control the transformation range for the unsupervised training phase described in 2.2. That transformation can be defined as the combination of four operations : translation, rotation, shearing and scaling. Several sets of hyperparameters ranges are considered in the experiments and are presented in table 1. Note that the "easy" parameter range D_0 includes a full range of rotation.

During the testing phase, we apply to the registered images I_{test} a random affine transformation T' with a given parameters range of table 1. As the ideal affine registration is the inverse of the random transform used to unregister the images, the following metrics are used for a transform T given by the network:

$$\text{Mat}_{\text{err}} = \|T \circ T' - Id\|^2 \qquad \text{and} \qquad \theta = \arccos \frac{tr(T \circ T')}{2}. \tag{6}$$

Robustness of the method is measured with θ_t defined as the percentage of images with a rotation error θ greater than t.

Implementation details The code is written in Pytorch/Monai. In our experiments, the network is implemented using the Monai class Regressor with six residual convolutional layers with respectively 8,16,32,64,128 and 256 channels. The convolution kernel sizes are 3x3 with the stride set to 1 for the first layer and set to 2 for the other layers. A PRELU is used after each convolutional layer.

Loss	Mat_{err}	θ_{90}	θ_{45}
L_{equiv}	1.68 ± 0.14	50.49	100.00
L_{equiv} $+L_{\text{size}}$ $+L_{\text{anis}}$	2.18 ± 0.12	35.36	100.00
L_{equiv} $+L_{\text{size}}$ $+L_{\text{anis}}$ +MI	2.56 ± 0.12	99.84	100.00
L_{equiv} $+L_{\text{ref}}$	1.66 ± 0.13	33.39	100.00
L_{equiv} $+L_{\text{ref}}$ $+L_{\text{size}}$	$\mathbf{0.45 \pm 0.29}$	**0.00**	**0.00**
L_{equiv} $+L_{\text{ref}}$ $+L_{\text{size}}$ $+L_{\text{anis}}$	$\mathbf{0.45 \pm 0.27}$	**0.00**	**0.00**
L_{equiv} $+L_{\text{ref}}$ $+L_{\text{size}}$ $+L_{\text{anis}}$ +MI	$\mathbf{0.39 \pm 0.26}$	**0.00**	**0.00**
MI	3.63 ± 1.08	50.55	100.00
FSL flirt - MI	2.58 ± 1.21	49.22	98.08
supervized - L_{ref}	$\mathbf{0.32 \pm 0.22}$	**0.00**	**0.00**

Table 2: Ablation study on the losses of our unsupervised one step method and comparison to state of the art on the D_0 parameters range.

The output consists in the six coefficients of the affine transformation matrix. Note that the training is not sensitive to the hyperparameters K of L_{reg}: a value of $K = 4$ is sufficient to train the network and a scale change of 4 is already a huge global scale change between two subjects. Two matrix norms are evaluated for L_{ref}: the Froboenius norm (Frob) and the norm $\|A\| = \sum_i \|Ax_i\|$ where x_i are uniformely sampled in the image (denoted as Grid). A 3x3 grid is used here.

3.2 Results & Discussion

Ablation study using D_0 for the one step method An ablation study was carried out to assess the importance of each loss in our one step method. Evaluation metrics for different setups are presented in table 2 using D_0 as the affine parameters range for train and test. One can see that the combination of our three losses L_{ref}, L_{reg} and L_{equiv} is essential for our method to be robust. If one of these losses is missing, the rotation error θ_{45} will be higher than 93% and the use of MI does not help. When our three losses are used, θ_{45} drops to zero. One can also notice the very low value of Mat_{err} and that adding MI to our three losses does help here for the D_0 range of affine parameters. As a upper bound on the performance that can be achieved, a comparison was made with a model (supervized - L_{ref} trained with the L_{ref} loss used for all images (all registered) in the training dataset . Robustness error is null for this model. One can notice that Mat_{err} is the lowest for this model but our unsupervised model achieves a Mat_{err} error that is not too high compared to this upper bound model. Our results have also been compared to an unsupervised model trained only with the MI loss between the warped image and I_{ref} and to FSL - flirt [6]. As D_0 includes the full range of possible rotations, these two methods have a large rotation error. Indeed, due to the symmetry in brain images, the MI loss is not able to correctly handle large rotation. Despite this symmetry obstacle, our unsupervised method correctly register images with a full range of rotation.

Range		Mat_{err}	θ_{90}	θ_{45}
L_{ref}	L_{equiv}			
D_0	D_0	0.41 ± 0.27	0.00	0.00
D_{shear}	D_{shear}	1.29 ± 0.68	0.00	0.00
D_{trans}	D_{trans}	2.96 ± 1.70	49.51	84.70
D_{scale}	D_{scale}	3.27 ± 0.58	83.39	83.39
D_1	D_0	20.33 ± 17.11	7.89	44.08
D_0	D_1	11.71 ± 6.06	46.71	86.68
D_1	D_1	12.08 ± 5.70	41.78	81.10

Table 3: Influence of the affine parameter range of the B matrix on our method

Method	Mat_{err}	θ_{90}	θ_{45}
One step	12.08 ± 5.70	41.78	81.10
Two steps - MI	10.28 ± 5.59	38.98	56.09
Two steps - Frob.	**1.88 ± 0.90**	0.49	8.55
Two steps - Grid	3.69 ± 3.54	**0.00**	**1.32**
Two step - Frob - w. occ	4.55 ± 4.68	3.78	18.91
Two step - Grid - w. occ	**3.79 ± 3.67**	**0.00**	**1.64**

Table 4: Large affine transform registration: comparison with D_1 parameter range of our unsupervised one step method and our two steps method with differents losses in the second step, with and without occlusions.

Robustsness to large affine parameters ranges To evaluate the ability of our unsupervised method to register images with large affine transforms, our model with our three losses (without MI) was trained with several affine parameter ranges for L_{ref} and L_{equiv}. Results are presented in table 3. As already noticed, despite D_0 includes the full range of possible rotations, our method is able to correctly find the correct transform with no rotation error and a low Mat_{err}. Our method is also robust to large shear although Mat_{err} is higher in this case. One can also note that our method is more robust to an increase in range for the L_{ref} loss than for the L_{equiv} loss. This last point could be expected as L_{ref} is a supervised loss (but with only a single image). As a conclusion, although it is remarkable that a regular training of the network with L_{equiv}, L_{ref} and L_{reg} enables to obtain a robust template registration for large shear and rotation, it is not sufficient for general large affine transform and the training procedure of section 2.2 should be used.

Large transform registration with the two steps training In table 4, a comparison of our one step and our two steps method is presented. The D_0 parameter range was used for the first step, D_1 was used for the second step and the evaluation. For the second step, the L_{ref} loss was implemented using either the Frob. or the Grid norm. A two steps method with MI in the second step was

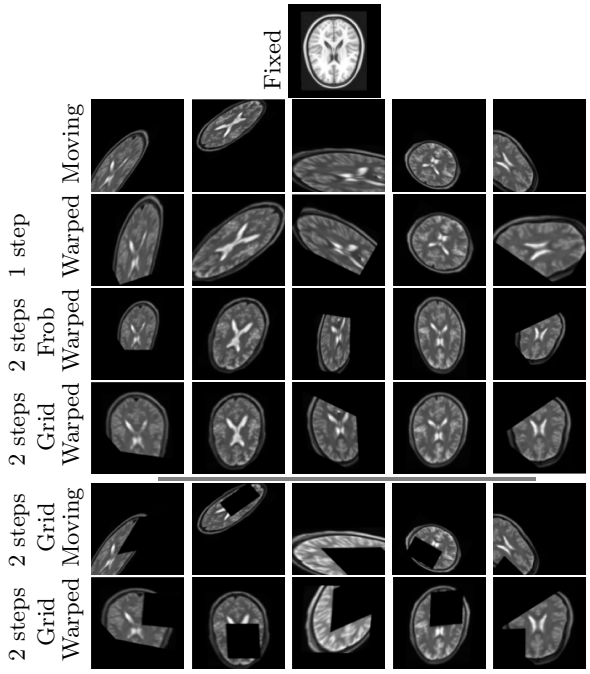


Fig. 2: Large affine transform registration: comparison with D_1 parameter range of our unsupervised one step method and our two steps method with differents losses in the second step. Results with / without occlusions.

also evaluated. Large occlusions were also added during train and test to test the robustness to abnormalities.

One can see that both "one step" and "two steps with MI" are unable to correctly register the images for the D_1 range of affine parameters. In contrast, the robustness error drops considerably for our two steps approaches with both Frob. or Grid matrix norm. Note that although Mat_{err} is lower for Frob, both θ_{45} and θ_{90} are lower with Grid. Superiority of Grid for the training is confirmed by visual inspection: one can clearly see in fig. 2, that "two steps - Grid" is the only method that correctly realign the images in the reference coordinate system despite the large initial misalignment, even for the extreme case where a large part of the image is cropped. Note that the superiority of Grid over Frob was also reported in [3] for homography estimation. Finally, one can see using both the quantitative metrics and the visual inspection, that the two steps procedure with the Grid norm is able to correctly register images when both large abnormalities are present on the image and the affine transform is very large.

4 Conclusion

In this paper, we proposed an unsupervised image registration method to train a network that directly outputs the affine transform of an image to an atlas. Our method does not rely on dissimilarity metrics but on three losses that enforce prior on the registration tasks: equivariance, invertibility of the output and the positioning of a unique given template. This simple method is able to robustly register when full range of rotation and large shear are present. For large translation, large scale change, a two steps training procedure is used to enlarge the range of affine matrixes that can be estimated to the most extreme cases. Numerical evaluation shows the robustness of our approach even in the presence of large crops or occlusions. In a future work, we plan to extend our approach to the deformable registration case.

Acknowledgments This work was supported by the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, by the "Projet Emergence" CNRS-INS2I APIDIFF, by the INSA BQR SALVE and by the France Life Imaging network (ANR-11-INBS-0006). Experiments were carried out using HPC resources from GENCI-IDRIS (AD011012544/AD011012589).

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
2. Czolbe, S., Pegios, P., Krause, O., Feragen, A.: Semantic similarity metrics for image registration. Medical Image Analysis **87**, 102830 (2023). https://doi.org/https://doi.org/10.1016/j.media.2023.102830
3. DeTone, D., Malisiewicz, T., Rabinovich, A.: Deep image homography estimation. In: RSS Workshop: Are the Sceptics Right? Limits and Potentials of Deep Learning in Robotics (06 2016)
4. Evan, M.Y., Wang, A.Q., Dalca, A.V., Sabuncu, M.R.: Keymorph: Robust multimodal affine registration via unsupervised keypoint detection. In: Medical Imaging with Deep Learning (2022)
5. Hill, D., Batchelor, P., Holden, M., Hawkes, D.: Medical image registration. Physics in medicine and biology **46**, R1–45 (04 2001). https://doi.org/10.1088/0031-9155/46/3/201
6. Jenkinson, M., Bannister, P., Brady, M., Smith, S.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage **17**(2), 825–841 (2002)
7. Li, H., Fan, Y.: Non-rigid image registration using self-supervised fully convolutional networks without training data. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 1075–1078. IEEE (2018)
8. Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., Comaniciu, D.: An artificial agent for robust image registration. Proceedings of the AAAI Conference on Artificial Intelligence **31**(1) (Feb 2017). https://doi.org/10.1609/aaai.v31i1.11230

9. Mattes, D., Haynor, D.R., Vesselle, H., Lewellyn, T.K., Eubank, W.: Nonrigid multimodality image registration. In: Medical imaging 2001: image processing. vol. 4322, pp. 1609–1620. Spie (2001)
10. Miao, S., Wang, Z.J., Liao, R.: Real-time 2d/3d registration via cnn regression (2016)
11. Rohé, M.M., Datar, M., Heimann, T., Sermesant, M., Pennec, X.: Svf-net: learning deformable image registration using shape matching. In: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20. pp. 266–274. Springer (2017)
12. Siebert, H., Hansen, L., Heinrich, M.P.: Learning a metric for multimodal medical image registration without supervision based on cycle constraints. Sensors **22**(3), 1107 (2022)
13. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE Transactions on Medical Imaging **32**(7), 1153–1190 (2013). https://doi.org/10.1109/TMI.2013.2265603
14. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Medical Image Analysis **52**, 128–143 (feb 2019)