A Deep Metric for Multimodal Registration

Martin Simonovsky[1(✉)], Benjamín Gutiérrez-Becker[2], Diana Mateus[2], Nassir Navab[2], and Nikos Komodakis[1]

[1] Imagine, Université Paris Est/École des Ponts ParisTech,
Champs-sur-Marne, France
{martin.simonovsky,nikos.komodakis}@enpc.fr

[2] Computer Aided Medical Procedures, Technische Universität München,
Munich, Germany
gutierrez.becker@tum.de, {mateus,navab}@in.tum.de

Abstract. Multimodal registration is a challenging problem due the high variability of tissue appearance under different imaging modalities. The crucial component here is the choice of the right similarity measure. We make a step towards a general learning-based solution that can be adapted to specific situations and present a metric based on a convolutional neural network. Our network can be trained from scratch even from a few aligned image pairs. The metric is validated on intersubject deformable registration on a dataset different from the one used for training, demonstrating good generalization. In this task, we outperform mutual information by a significant margin.

1 Introduction

Multimodal registration is a very challenging problem commonly faced during image-guided interventions and data fusion [12]. The main difficulty of the multimodal registration task comes from the great variability of tissue or organ appearance when imaged by different physical principles, which translates in the lack of a general rule to compare such images. Therefore, efforts to tackle this problem focus mainly on the design of multimodal similarity metrics.

Recent works have explored the use of supervised methods to learn similarity metrics from a set of aligned examples [2,7,10], showing potential to outperform hand-crafted metrics in particular applications. However, a general method to learn similarity between any two modalities calls for higher capacity models.

Inspired by their success in computer vision, we propose to learn such general similarity metric based on Convolutional Neural Networks (CNNs). The problem is modelled as a classification task, where the goal is to discriminate between aligned and misaligned patches from different modalities. To the best of our knowledge, this is the first time that CNNs are used in the context of multimodal medical image registration.

The ability of our metric to obtain reliable registrations is demonstrated on the ALBERTs database of neonatal images [5], where we outperform Mutual Information [9]. Importantly, we train on a separate dataset (IXI database of

S. Ourselin et al. (Eds.): MICCAI 2016, Part III, LNCS 9902, pp. 10–18, 2016.
DOI: 10.1007/978-3-319-46726-9_2

adults [1]), demonstrating the capability to generalize to data acquired with different scanners and with demographic differences in the subjects. We also show that our method is able to learn reliable multimodal similarities even with a small training set, as is often the case in medical imaging applications.

1.1 Related Work

The idea of using supervised learning to build a similarity metric for multimodal images has been explored in a number of works. On one side, there are probabilistic approaches which rely on modelling the joint-image distribution. For instance, Guetter *et al.* propose a generative method based on Kullback-Leibler Divergence [6]. Our work is closer to the discriminative concept proposed by Lee *et al.* [7] and Michel *et al.* [10], where the problem of learning a similarity metric is posed as binary classification. Here the goal is to discriminate between aligned and misaligned patches given pairs of aligned images. Lee *et al.* propose the use of a Structured Support Vector Machine while Michel *et al.* use a method based on Adaboost. Different to these approaches we rely on CNN as our learning method of choice as the suitable set of characteristics for each type of modality combinations can be directly learned from the training data.

The power of CNNs to capture complex relationships between multimodal medical images has been shown in the problem of modality synthesis [11], where CNNs are used to map MRI-T2 images to MRI-T1 images using jointly the appearance of a small patch together with its localization. Our work is arguably most similar to the approach of Cheng *et al.* [3] who train a multilayer fully-connected network pretrained with autoencoder for estimating similarity of 2D CT-MR patch pairs. Our network is a CNN, which enables us to scale to 3D due to weight sharing and train from scratch. Moreover, we evaluate our metric on the actual task of registration, unlike Cheng *et al.*

2 Method

Image registration is the task of estimating the best spatial transformation $\mathcal{T} : \Omega_f \mapsto \mathbb{R}^d$ between a *fixed image* $I_f : \Omega_f \subset \mathbb{R}^d \mapsto \mathbb{R}$ and a *moving image* $I_m : \Omega_m \subset \mathbb{R}^d \mapsto \mathbb{R}$. In our setting $d = 3$ and the images come each from a different modality. The problem is often solved by minimizing the energy

$$E(\theta) = M(I_f, I_m(\mathcal{T}(\theta))) + R(\mathcal{T}(\theta)) \tag{1}$$

where the first term M is a metric quantifying the cost of the alignment by transformation $\mathcal{T}$ parameterized by θ and the second term R is a regularization constraining the mapping. We denote the moving image resampled into Ω_f by $\mathcal{T}$ as the *warped image* $I'_m = I_m(\mathcal{T}(\theta)) : \Omega_f \subset \mathbb{R}^d \mapsto \mathbb{R}$. The minimization is commonly solved in a continuous or discrete optimization framework [12], depending on the nature of θ.

In this work we explore formulating M as a convolutional neural network. To this end we rely on network $N(P_f, P_m)$ which outputs a scalar value estimating the dissimilarity between two image patches $P_f \subset I_f$ and $P_m \subset I'_m$ of the same size. Its incorporation into a continuous optimization framework is explained in Subsect. 2.1. The architecture and training of N is described in Subsect. 2.2.

2.1 Continuous Optimization

Continuous optimization methods iteratively update parameters θ based on the gradient of the objective function $E(\theta)$. We restrict ourselves to first-order methods and use gradient descent in particular. Our metric is defined to aggregate local patch comparisons as

$$M(I_f, I'_m) = \sum_{P \in \mathcal{P}} N(I_f(P), I'_m(P)) \tag{2}$$

where $\mathcal{P}$ is the set of patch domains $P \subset \Omega_f$ sampled on a dense uniform grid with significant overlaps.

Its gradient $\nabla M(\theta)$, which is required for $\nabla E(\theta)$, can be computed by applying chain rule as follows:

$$\frac{\partial \sum_{P \in \mathcal{P}} N(I_f(P), I'_m(P))}{\partial \theta} = \sum_{\mathbf{x} \in \Omega_f} \sum_{P \in \mathcal{P}_\mathbf{x}} \frac{\partial N(I_f(P), I'_m(P))}{\partial I'_m(\mathbf{x})} \frac{\partial I'_m(\mathbf{x})}{\partial \theta} \tag{3}$$

$$= \sum_{\mathbf{x} \in \Omega_f} \sum_{P \in \mathcal{P}_\mathbf{x}} \frac{\partial N(I_f(P), I'_m(P))}{\partial I'_m(\mathbf{x})} \frac{\partial I_m(\mathcal{T}(\theta, \mathbf{x}))}{\partial \mathcal{T}(\theta, \mathbf{x})} \frac{\partial \mathcal{T}(\theta, \mathbf{x})}{\partial \theta} \tag{4}$$

$$= \sum_{\mathbf{x} \in \Omega_f} \frac{\partial N(I_f, I'_m)}{\partial I'_m(\mathbf{x})} \nabla I_m(\mathcal{T}(\theta, \mathbf{x})) J_{\mathcal{T}}(\mathbf{x}) \tag{5}$$

Equation (3) shows that the derivative of N w.r.t. the intensity of an input pixel $\mathbf{x}$ depends on all patches containing it, denoted as $\mathcal{P}_\mathbf{x}$. Thus, high overlap of neighboring patches leads to smoother, more stable derivatives. We found that registration quality drops considerably unless the grid stride s of $\mathcal{P}$ is small. On the other hand, subsampling Ω_f to obtain a sparser set of samples $\mathbf{x}$ has a minor impact on performance.

In the transition from Eqs. (4) to (5), patch-wise evaluation of N is replaced by fully convolutional evaluation over the whole domain Ω_f. This makes the computation very efficient, as results in intermediate network layers can be shared among neighboring patches [8].

Ultimately, the contribution of each pixel $\mathbf{x}$ to $\nabla M(\theta)$ is a product of three terms, *c.f.* Eq. (5): the derivative $\partial N / \partial I'_m(\mathbf{x})$ of the estimated dissimilarity of patches around $\mathbf{x}$ w.r.t. its intensity in the warped image, which can be readily computed by standard backpropagation, the gradient of the moving image ∇I_m, which can be precomputed, and the local Jacobian matrix $J_{\mathcal{T}}$ of transformation $\mathcal{T}$. Note that the choice of a particular transformation type is decoupled from the network, therefore a single network will work with any transformation.

Computing one iteration thus requires resampling of the moving image and one forward and one backward pass in the network. All operations can be efficiently computed on a GPU.

2.2 Network Architecture and Training

Architecture. A feed-forward convolutional neural network N is used to estimate the dissimilarity of two cubic patches of the same size of $p \times p \times p$ pixels. The architecture is based on recent works on learning to compare patches, notably the 2-channel network of Zagoruyko and Komodakis [13]. The two patches are considered as a 2-channel 3D image (each channel represents a different modality), which is fed to the first layer of the network. The network consists of a series of volumetric convolutional layers with ReLU non-linearities finalized by a convolutional layer without any non-linearity, which produces a scalar score.

To gradually subsample the spatial domain within the network and increase spatial context (perceptive field), we prefer convolutions with non-unit output stride to pooling used in [13], as it has led to better performance. We hypothesize that too much spatial invariance might be detrimental in our case of learning cross-modal identity, unlike aiming for robustness to distortions such as perspective deformation. The product of convolutional strides determines the overall network stride s used in the fully-convolutional mode.

The 2-channel architecture is powerful as it considers both patches jointly from the beginning. However, its evaluation does not exploit the fact that the fixed image I_f does not change during optimization and its deep representation could be precomputed in the form of descriptors and cached. We have therefore experimented on architectures with two independent input branches, such as the pseudo-siamese network in [13]. Unfortunately, we have observed consistent decrease in registration performance.

Training. We suppose to have a set of k aligned pairs of training images $\{(A_j, B_j)\}_{j=1}^{k}$ with $A_j, B_j : \Omega_j \subset \mathbb{R}^d \mapsto \mathbb{R}$. We sample transformations $\mathcal{T}_{i,A_j}$, $\mathcal{T}_{i,B_j} : \Omega_j \mapsto \Omega_j$ for j-th image pair for data augmentation by varying position, scale, rotation, and mirroring. Patch pairs $X_i = (A_j(\mathcal{T}_{i,A_j}(P)), B_j(\mathcal{T}_{i,B_j}(P)))$ with fixed-size domain P are used for training the network. Sample X_i is defined to be positive (labeled $y_i = -1$) if $\mathcal{T}_{i,A_j} = \mathcal{T}_{i,B_j}$ and negative ($y_i = 1$) otherwise. Positive and negative samples are mined with equal probability. Imposing restrictions on negatives (such as minimum or maximum overlap of source patch domains) or on patch content (such as minimum contrast [7]) were experimentally shown detrimental to the registration quality.

The network is trained to classify training samples X_i by minimizing hinge loss $L = \sum_i \max(0, 1 - y_i N(X_i))$, which we found to perform better than cross-entropy. We observed that softmax leads to overly flat gradients in continuous optimization, as shown in the bottom plots in Fig. 2. SGD with learning rate 0.01, momentum 0.9 and batch size 128 is used to optimize the network.

Instead of preparing a fixed dataset of patches like in [3], we sample X_i online. This, together with the augmentations described above, allows us to feed the network with practically unlimited amount of training data. Even for small k we observed no overfitting in learning (see also Subsect. 3.2).

Implementation. We use Torch with cuDNN library for deep learning, elastix for GPU-based image resampling, and ITK for registration[1] Our network has 5 layers, 2M parameters, patch size $p = 17$, and stride $s = 4$. We plan to open source our implementation and the trained network.

3 Experiments and Results

We evaluate the effectiveness of the learned metric in registration experiments on a set of clinical brain images in Subsect. 3.1 and conduct further experiments to demonstrate its interesting properties in Subsects. 3.2 and 3.3.

3.1 Deformable Registration of Neonatal Brain MRI Images

Datasets. We conducted intersubject deformable registration experiments on a set of neonatal brain image volumes taken from the publicly available brain atlases ALBERTs [5]. This database consists of T1 and T2-weighted MRI scans of 20 newborns. Each T1-T2 pair is aligned and annotated with a segmentation map of 50 anatomical regions, which allows us to evaluate registration quality in terms of overlap measures; we compute average Dice and Jaccard coefficients.

To make the experiment challenging and demonstrate good generalization of our learned metric (denoted CNN), we train on IXI [1], a completely independent dataset of adult brain images. Let us remark that there are structural differences between the brains of neonates and adults. The dataset contains about 600 approximately aligned T1–T2 image pairs and we use $k = 557$ for training and the rest for validation, although in Subsect. 3.2 we demonstrate that much less is actually needed. Image intensities in both datasets are normalized to $[0, 1]$.

Baseline. Our baseline is mutual information (MI) [9], the standard metric for multimodal registration. We observed that MI perform better when image domains are restricted to the head region, thus we use a fixed intensity threshold of 0.01 for masking the background and denote this variant MI+M. Such masking made nearly no difference to our metric. Unfortunately, we could not compare to other learning-based metrics [7,10] as their implementation was not available.

Protocol. We test on 18 subjects in ALBERTs and perform 68 intersubject registrations, half of them aligning T1 to T2 and half of them the other way round. We reserve the remaining 2 subjects for validating registration parameters and model selection. Both metrics are evaluated in exactly the same registration pipeline with the same transformation model and optimizer. The pipeline consists of multiresolution similarity transform registration followed by multiresolution B-spline registration (2 scales, 1000 control points on the fine scale, 200 k image sampling points), optimized by gradient descent with regular step and 500 iterations per scale. MI is used with 75 histogram bins (validated optimum). An explicit regularization term R in Eq. (1) was used neither for MI

[1] www.torch.ch, developer.nvidia.com/cudnn, elastix.isi.uu.nl, www.itk.org.

Table 1. Overlap scores (mean ± SD) after registration using the proposed metric (CNN) and mutual information with (MI+M) or without masking (MI)

	MI+M	MI	CNN $k = 557$	CNN $k = 11$	CNN $k = 6$	CNN $k = 3$
Dice	0.665 ± 0.096	0.497 ± 0.180	0.703 ± 0.037	0.704 ± 0.037	0.701 ± 0.040	0.675 ± 0.093
Jaccard	0.519 ± 0.091	0.369 ± 0.151	0.555 ± 0.041	0.556 ± 0.041	0.554 ± 0.044	0.527 ± 0.081

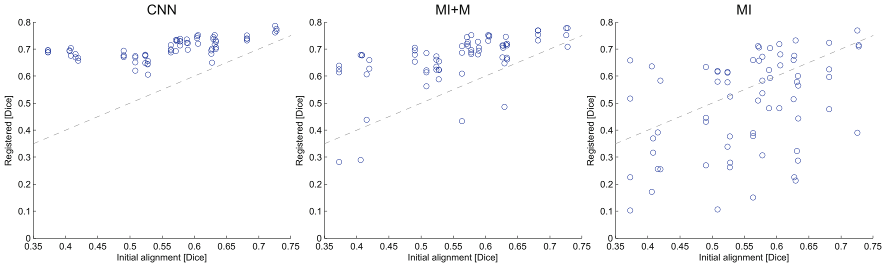

Fig. 1. Improvement in average Dice score due to registration using the proposed metric (CNN) and mutual information with (MI+M) or without masking (MI). Each data point represents a registration run. Dashed line denotes identity transformation.

nor for CNN. Instead, we regularize implicitly by the design of the pipeline and the choice of its hyperparameters.

Results. The results are listed in Table 1 and demonstrate statistically significant improvement of registration quality due to CNN by about 4 points in both coefficients (as by one-sided t-test with significance $\alpha = 0.01$). Figure 1 exhibits scatter plots of initial and final Dice scores for each registration run (Jaccard scores follow similar trend). We can see that while CNN has improved on the alignment in all runs, this is not the case for MI+M and especially MI, showing rather low precision. The highest accuracies achieved by both methods are rather similar (up to 0.8) and seem nearly independent on the initial level of misalignment. Furthermore, the registration using CNN is only about 2x slower than using MI (on Nvidia Titan Black), the difference mostly due to expensive resampling of moving image.

3.2 Influence of Training Set Size

The huge number of aligned volumes in IXI dataset is rather exceptional in medical domain. We are therefore interested in how much we can decrease the training set size k without noticeable impact on the quality. To this end, we train networks with only $k = 11$, 6, and 3 random image pairs under the same setting as above. Table 1 shows that even with little training data the results are very good and only for $k = 3$ our metric does not significantly outperform MI+M. On one hand, this suggests that our online sampling and data augmentation methodology works well. On the other hand, either the inherent variability in the dataset is very low (especially compared to natural image recognition problems, where more data typically improves performance) or our network is not able to exploit it. We expect that the amount of necessary data will be higher for more challenging modalities, such as ultrasound.

3.3 Plausibility of Metric and Its Derivatives

To investigate the behavior of metric value and its actual derivatives used for continuous optimization, we visualize these quantities by manually perturbing a single parameter of a transformation initialized to identity on an aligned validation image pair in IXI. Figure 2 suggests that the metric behaves reasonably as its curves are smooth with the correct local minima. The analytic derivatives, as in Eq. (5), have the correct sign over a large range, albeit their magnitude is slightly noisy. Nevertheless, this was shown not to prevent the metric from obtaining good registration results.

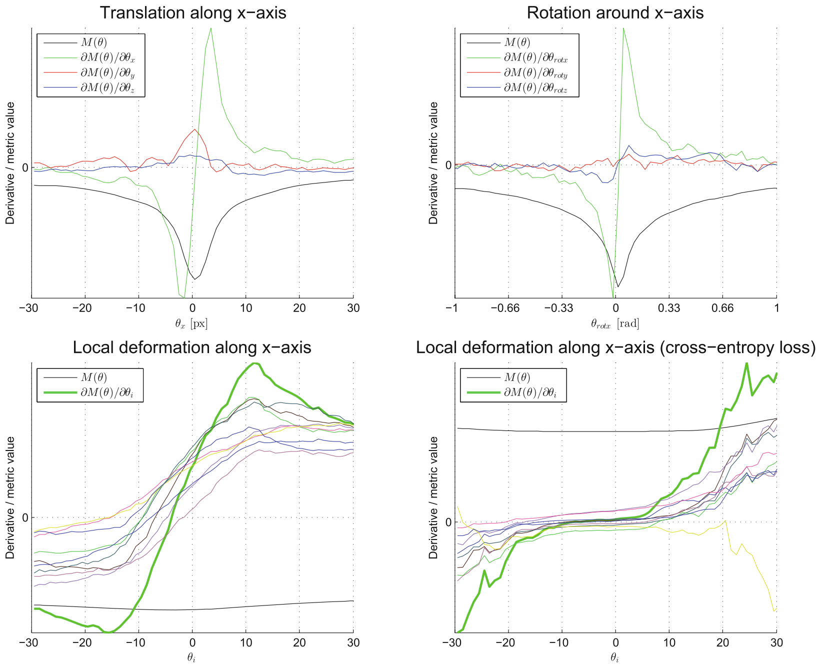

Fig. 2. The impact of perturbation of a single parameter in Euclidean transform (top) and B-spline transform (bottom) on metric value M and its derivatives as per Eq. (5). The curve of M is not up to scale. Curves without legend correspond to other parameters strongly affected due to overlapping patches.

4 Conclusion

We have presented a similarity metric for multimodal 3D image registration based on a convolutional neural network. The network can be trained from scratch even from a few aligned image pairs, mostly due to our data sampling scheme. We have described the incorporation of this metric into first-order continuous optimization frameworks. The experimetal evaluation was performed on the task of intersubject T1–T2 deformable registration on a dataset different from the one used for training, demonstrating good generalization. In this task, we outperform mutual information by a significant margin.

We envision incorporating our network into a discrete optimization framework as an easy extension. In a MRF-based formulation, the local alignment cost is expressed by unary potentials over nodes in a graph [4]. In particular, a unary potential $g_n(\mathbf{u}_n)$ related to the cost of assigning a label/translation $\mathbf{u}_n$ to node n might be defined as $g_n(\mathbf{u}_n) = N(I_f(P_n), I_m(\mathcal{T}(\theta, P_n) + \mathbf{u}_n))$, where $P_n \subset \Omega_f$ is a patch domain centered at the control point of transformation $\mathcal{T}$ corresponding to node n. As such an optimization is derivative-free, only the forward pass in the network would be necessary.

We also plan to apply our method to more modalities, such as ultrasound.

Acknowledgments. We gratefully acknowledge NVIDIA Corporation for the donated GPU used in this research. ALBERTs atlases are copyrighted by Imperial College of Science, Technology and Medicine and Ioannis S. Gousias 2013. B. Gutiérrez-Becker thanks the financial support of CONACYT and the DAAD.

References

1. IXI - Information eXtraction from images. www.brain-development.org
2. Cao, T., Jojic, V., Modla, S., Powell, D., Czymmek, K., Niethammer, M.: Robust multimodal dictionary learning. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 259–266. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40811-3_33
3. Cheng, X., Zhang, L., Zheng, Y.: Deep similarity learning for multimodal medical images. In: Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization, pp. 1–5 (2015)
4. Glocker, B., Sotiras, A., Komodakis, N., Paragios, N.: Deformable medical image registration: setting the state of the art with discrete methods. Ann. Rev. Biomed. Eng. **13**, 219–244 (2011)
5. Gousias, I., Edwards, A., Rutherford, M., Counsell, S., Hajnal, J., Rueckert, D., Hammers, A.: Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. Neuroimage **62**, 1499–1509 (2012). www.brain-development.org
6. Guetter, C., Xu, C., Sauer, F., Hornegger, J.: Learning based non-rigid multi-modal image registration using kullback-leibler divergence. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3750, pp. 255–262. Springer, Heidelberg (2005). doi:10.1007/11566489_32

7. Lee, D., Hofmann, M., Steinke, F., Altun, Y., Cahill, N., Scholkopf, B.: Learning similarity measure for multi-modal 3D image registration. In: CVPR, pp. 186–193 (2009)
8. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
9. Mattes, D., Haynor, D.R., Vesselle, H., Lewellyn, T.K., Eubank, W.: Nonrigid multimodality image registration. In: SPIE, vol. 4322, pp. 1609–1620 (2001)
10. Michel, F., Bronstein, M., Bronstein, A., Paragios, N.: Boosted metric learning for 3D multi-modal deformable registration. In: ISBI, pp. 1209–1214 (2011)
11. Nguyen, H., Zhou, K., Vemulapalli, R.: Cross-domain synthesis of medical images using efficient location-sensitive deep network. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 677–684. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24553-9_83
12. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: a survey. TMI **32**(7), 1153–1190 (2013)
13. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: CVPR, pp. 4353–4361 (2015)